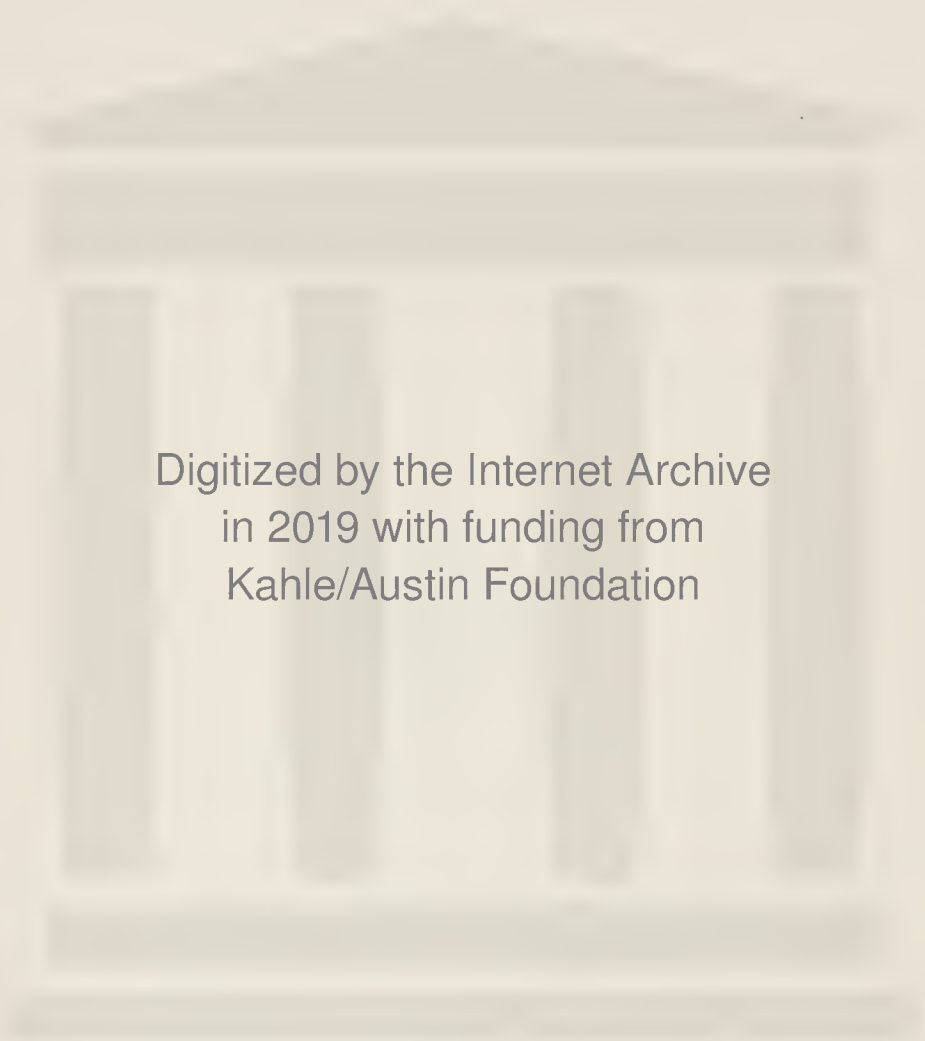


NUNC COGNOSCO EX PARTE



TRENT UNIVERSITY
LIBRARY



Digitized by the Internet Archive
in 2019 with funding from
Kahle/Austin Foundation

NUMERICAL ANALYSIS

NUMERICAL ANALYSIS

BY

D. R. HARTREE, F.R.S.

*Plummer Professor of Mathematical Physics
University of Cambridge*

SECOND EDITION

OXFORD
AT THE CLARENDON PRESS

Oxford University Press, Amen House, London E.C.4

GLASGOW NEW YORK TORONTO MELBOURNE WELLINGTON
BOMBAY CALCUTTA MADRAS KARACHI KUALA LUMPUR
CAPE TOWN IBADAN NAIROBI ACCRA

© *Oxford University Press 1958*

FIRST EDITION 1952
REPRINTED LITHOGRAPHICALLY IN GREAT BRITAIN
FROM SHEETS OF THE FIRST EDITION 1954
SECOND EDITION 1958
REPRINTED LITHOGRAPHICALLY 1961

ONULP

PREFACE TO THE SECOND EDITION

IN revising the text of this book for a second edition, it has been my intention to preserve its character as an introduction to numerical analysis for those who want to know about numerical methods for the purpose of applying them in practice, when actual numbers take the place of the literal symbols of an algebraic formula.

The main change from the first edition is in Chapter VII on the numerical integration of differential equations, which has been largely rearranged and somewhat extended, particularly in the treatment of equations with two-point boundary conditions. I have also included a section on Whittaker's 'cardinal function' in the theory of interpolation, and an account of a method, due to Wilkes, of handling the Choleski method of matrix factorization which, in my view, converts this from a method for specialists (as it seemed to me when the first edition was written) to a practicable method for the occasional user. The treatment of quadrature formulae of the Gauss type has been extended, and a number of other sections have been added or revised.

When the first edition was written, little had been published on programming for automatic digital computers, and it seemed advisable to include an introduction to this subject with some simple examples; these required the use of a particular programming system, which had to be explained. Since then, much more has been published on the subject (see references on p. 284), and I have consequently excised those sections concerned with details of programming.

I wish to thank various correspondents, particularly Mr. R. E. Beard, Mr. D. R. Bland, Mr. G. A. Erskine, and Mr. M. Fine, who have written to draw my attention to points in the first edition requiring correction or modification. It is a pleasure to renew my thanks to the staff of the Clarendon Press for their co-operation.

D. R. H.

Cavendish Laboratory
Cambridge
October 1957

PREFACE TO THE FIRST EDITION

THIS book is based on a course of lectures on Numerical Analysis which I have given in the Mathematical Laboratory of the University of Cambridge for several years. It is intended to be introductory, in the sense that no previous knowledge of the theory and practice of systematic numerical work is assumed, but it is not 'elementary' in the sense of using only school mathematics. It assumes familiarity with the calculus up to Taylor's theorem and partial derivatives, acquaintance with differential equations and, in the chapter on linear simultaneous algebraic equations, with some of the simpler properties of matrices. But in all these cases what is wanted is mainly an understanding of the ideas involved rather than technical facility in manipulating algebraical or analytical expressions.

I have deliberately tried to restrict the algebraical and analytical work to the treatment of those methods which are useful in practice when numbers are substituted for the literal symbols of a general treatment, and to avoid developments which are of purely formal interest. Such developments may be elegant mathematics, or may make the formal presentation more complete, but they are not contributions to numerical analysis, and are distracting rather than helpful to the reader who wants practical information about what calculations to make to obtain the results he requires, and how to carry them out. For a similar reason, I have tried to give prominence to the importance of checking in numerical work. Mistakes can occur in such work, and it follows that some process of checking is necessary to ensure that any results obtained are not vitiated by undetected mistakes. A treatment of numerical methods which does not pay some attention to this aspect of the subject seems to me to be quite unrealistic. In some worked examples I have deliberately introduced mistakes in order to show how, by suitable checking procedure, they can be detected, diagnosed, and corrected.

For similar practical reasons I have deliberately omitted some examples of numerical work which seem to have become almost classical, for example the evaluation of an approximation to $\frac{1}{4}\pi$ by application of Gregory's quadrature formula to the integral $\int_0^1 dx/(1+x^2)$. This particular calculation I regard as an example of how *not* to do numerical work; not because the method is wrong, but because it is not the most suitable for obtaining the required result (see § 6.51); use of it is therefore

an example of bad practice, and should not be presented as if it were an example of satisfactory numerical procedure. Much could be written on how *not* to do numerical calculations. I have mentioned in the text some procedures which should, in general, be avoided, mostly because they are not the most suitable for obtaining the results sought, but some because they are dangerous in that, if used without precautions, they may give wrong results. I have refrained from giving numerical illustrations of the dangers of such methods, except in one case (§ 5.81) in which it seemed advisable to give a warning, by means of a horrid example, of the dangers of a method for inverse interpolation which might at first sight appear attractive, and which has in fact been given in print as a usable method without any mention of the dangers.

An introductory treatment such as that of this book cannot cover the subject completely; several of the chapters from Chapter IV onward could well be expanded to form a volume each. In particular, this book hardly begins to touch the needs of the specialist or research worker in the subject of numerical analysis; its purpose is rather to give an introduction to the subject to workers in other fields of pure or applied science who may have to carry out calculations of a non-trivial magnitude. In such contexts the accuracy of the approximations or measurements underlying the calculation to be done do not usually justify working to a greater accuracy than six or seven figures, and often a smaller number will be adequate. On the other hand, the amount of calculation to be done to this accuracy may be considerable. In so far as the appropriate method for carrying out a calculation depends on the number of figures kept in it, emphasis is therefore placed on methods suitable for calculations of substantial extent to moderate accuracy, rather than for a few calculations to many figures. My personal experience of such work extends over 35 years, and most of this work has been concerned with calculations involving numerical approximations to some of the limiting processes of analysis, in particular integration (including the integration of differential equations). On the basis of this experience, I believe this to be one of the most important practical fields for the use of numerical methods, and have deliberately given it considerable prominence.

In the past few years there has been considerable development of high-speed automatic general-purpose digital calculating machines. This development has given practical importance to the study of the process of organizing calculations for these machines, a process usually referred to as 'programming'. This study can be regarded as a branch of numerical analysis, and it is one which is likely to grow in importance

as more of these machines become available. I have therefore included an introductory chapter on this subject (Chapter XII). There are several systems of programming, and for brevity in an introductory account such as this, and to avoid confusing the reader with a number of alternatives, it seems best to adopt one particular system in presenting illustrative examples. I have adopted the one with which I myself am most familiar, and which I think is one of the simplest to follow. But this must not be regarded as more than a means for illustrating in a simple form some of the general ideas involved in programming. For various reasons, some of which are mentioned in § 12.8, † numerical methods which are convenient for hand calculation with the assistance of a desk machine are not the most suitable for an automatic machine, and vice versa. But this book is intended to provide an introduction to numerical analysis for those who will mainly be concerned with methods suitable for hand calculation, and little or no reference is made to other methods some of which might be more suitable for automatic machines.

In the Bibliography I have included, as suggestions for further reading, some books and papers not referred to in the text, but I have not attempted to compile a complete bibliography of numerical analysis, or to give references to the history of the subject; the reader who is interested in early references should consult *The Calculus of Observations* by Whittaker and Robinson.

On matters concerning the use of desk calculating machines, I am conscious of a considerable debt to the late Dr. L. J. Comrie; the processes of §§ 2.25, 4.45, 4.46, and Example 7‡ I learnt from him, though whether he originated them I do not know; and there may well be other examples of his influence of which I am unconscious. Some of Dr. Comrie's long and varied experience in numerical work is incorporated in *Chambers's Six-figure Mathematical Tables*, but it is much to be regretted that he did not live to write a fuller work on the art of numerical calculation.

In the derivation of central-difference interpolation formulae, I have followed a treatment which I learnt from J. G. L. Michel, and in the examination of truncation errors of interpolation and integration formulae I have followed a treatment which I learnt from Professor W. E. Milne while I was serving as Acting Chief of the Institute for Numerical Analysis of the U.S. National Bureau of Standards.

I wish to express my thanks to Dr. J. Howlett, of the Computing Section, A.E.R.E., who read the first draft typescript of this book, and

† § 12.3 of the present edition.

‡ Example 5 of the present edition.

to Mr. A. S. Douglas, who read the proof sheets, for many valuable comments and suggestions, and for a number of corrections. Also I wish to thank Mr. P. Farmer for the photographs from which the drawings for Figs. 1, 2, and 3 were made, Mrs. Valerie Taylor for making the drawings themselves, my daughter for her skill in typing much of the text, and Dr. M. V. Wilkes, Director of the Mathematical Laboratory, Cambridge, for permission to avail myself of the services of Mr. Farmer, Mrs. Taylor, and my daughter, all members of the staff of the Laboratory. It is a pleasure also to thank those members of the staff of the Clarendon Press who have been concerned with the production of this book.

D. R. H.

Cavendish Laboratory
Cambridge
May 1952

CONTENTS

| | |
|--|----|
| CHAPTER I. INTRODUCTION | 1 |
| 1.1. What numerical analysis is about | 1 |
| 1.2. The main types of problems in numerical analysis | 4 |
| 1.3. Errors, mistakes, and checking | 5 |
| 1.4. Arrangement of work | 8 |
| 1.5. Accuracy and precision | 9 |
| CHAPTER II. THE TOOLS OF NUMERICAL WORK AND HOW TO USE THEM | 11 |
| 2.1. The main tools of numerical work | 11 |
| 2.2. Desk machines | 11 |
| 2.21. Addition and subtraction | 14 |
| 2.22. Transfer from accumulator to setting keys or levers | 15 |
| 2.23. Multiplication | 16 |
| 2.24. Division | 17 |
| 2.25. Other calculations | 19 |
| 2.26. Adding machines | 19 |
| 2.3. Mathematical tables | 20 |
| 2.31. Critical tables | 21 |
| 2.32. Auxiliary variables in tables | 22 |
| 2.4. Slide rule | 23 |
| 2.5. Graph paper | 24 |
| 2.6. Other machines | 24 |
| CHAPTER III. EVALUATION OF FORMULAE | 26 |
| 3.1. The significance of formulae in numerical work | 26 |
| 3.2. Evaluation of polynomials | 28 |
| 3.3. Evaluation of power series | 29 |
| 3.4. Kinds of formulae to avoid | 30 |
| 3.5. Evaluation of a function in the neighbourhood of a value of the argument at which it becomes indeterminate | 32 |
| CHAPTER IV. FINITE DIFFERENCES | 33 |
| 4.1. Functions of a continuous variable in numerical analysis | 33 |
| 4.2. Finite differences | 35 |
| 4.21. Notation for finite differences | 36 |
| 4.3. Finite differences in terms of function values | 38 |
| 4.4. Simple applications of differences | 39 |
| 4.41. Differences of a polynomial | 39 |
| 4.42. Building up polynomials | 41 |
| 4.43. Checking by differences | 43 |
| 4.44. Effect of rounding errors on differences | 46 |

| | |
|--|-----|
| 4.45. Direct evaluation of second differences | 47 |
| 4.46. Building up from second differences | 48 |
| 4.5. Differences and derivatives | 49 |
| 4.6. Finite difference operators | 50 |
| 4.7. Examples of the use of finite difference operators | 54 |
| 4.71. Derivatives in terms of differences | 54 |
| 4.72. Negative powers of (U/δ) | 55 |
| 4.73. $\delta^2 f$ in terms of f'' and its differences | 56 |
| 4.74. $\delta f_{\frac{1}{2}}$ symmetrically in terms of f' and its differences at x_0 and x_1 | 57 |
| 4.75. $\mu\delta f_0$ in terms of f' and its differences at $x = x_0$ | 57 |
| CHAPTER V. INTERPOLATION | 59 |
| 5.1. Linear and non-linear interpolation | 59 |
| 5.11. Linear interpolation | 60 |
| 5.2. Non-linear interpolation | 61 |
| 5.21. Half-way interpolation | 61 |
| 5.22. Newton's forward-difference formula | 63 |
| 5.3. Some expansions | 64 |
| 5.4. Everett's interpolation formula | 66 |
| 5.41. Bessel's interpolation formula | 67 |
| 5.42. Use of Bessel's and Everett's formulae | 69 |
| 5.43. Practical details in non-linear interpolation | 71 |
| 5.5. Lagrange's formula | 74 |
| 5.51. Special interpolation methods for particular functions | 76 |
| 5.6. Subtabulation | 77 |
| 5.61. End-figure method for subtabulation | 79 |
| 5.7. Interpolation of a function given at unequal intervals of the argument | 82 |
| 5.71. Evaluation of Lagrange's interpolation formula by a sequence of linear cross-means | 84 |
| 5.72. Divided differences | 86 |
| 5.8. Inverse interpolation | 89 |
| 5.81. How not to do inverse interpolation | 91 |
| 5.9. Truncation errors in interpolation formulae | 93 |
| 5.91. Whittaker's cardinal function | 93 |
| CHAPTER VI. INTEGRATION (QUADRATURE) AND DIFFERENTIATION | 97 |
| 6.1. Definite and indefinite integrals, and the integration of differential equations | 97 |
| 6.2. Integration formula in terms of integrand and its differences | 98 |
| 6.21. An alternative derivation | 99 |
| 6.22. Integration formula in terms of the integrand and the differences of its derivative | 100 |
| 6.23. Integration formula in terms of the integrand and its derivatives (Euler-Maclaurin formula) | 101 |

| | |
|---|-----|
| 6.3. Integration over more than one interval | 101 |
| 6.4. Evaluation of an integral as a function of its upper limit | 104 |
| 6.41. Change of interval length in an integration | 108 |
| 6.42. Integration in the neighbourhood of a singularity of the integrand | 110 |
| 6.43. Integration when the integrand increases 'exponentially' | 111 |
| 6.44. Two-fold integration | 112 |
| 6.5. Integrals between fixed limits | 113 |
| 6.51. Gregory's formula | 114 |
| 6.52. Integral in terms of function values | 114 |
| 6.53. Use of Simpson's or Weddle's rules | 115 |
| 6.54. Integrals of functions for which $f^{(2n+1)}(x) = 0$ at both ends of the range of integration | 115 |
| 6.55. Evaluation of a definite integral when the integrand has a singularity | 118 |
| 6.56. Definite integrals which are functions of a parameter | 118 |
| 6.6. Use of unequal intervals of the independent variables | 120 |
| 6.61. Gaussian integration formulae | 120 |
| 6.62. Gaussian formulae for $\int_0^{\infty} e^{-kx} p_{2n+1}(x) dx$ | 123 |
| 6.7. Numerical differentiation | 124 |
| 6.71. Differentiation formulae | 126 |
| 6.72. Graphical differentiation | 129 |
| 6.8. Errors of interpolation and integration formulae | 129 |
| 6.81. Use of formulae for the error | 132 |
| CHAPTER VII. INTEGRATION OF ORDINARY DIFFERENTIAL EQUATIONS | 134 |
| 7.1. Step-by-step methods | 134 |
| 7.11. One-point and two-point boundary conditions | 134 |
| 7.2. Second-order equation with first derivative absent | 135 |
| 7.21. Change of the interval of integration | 139 |
| 7.22. Variants of the method | 141 |
| 7.23. Numerov's method | 142 |
| 7.3. First-order differential equations | 143 |
| 7.31. Another method for a first-order equation | 146 |
| 7.32. First-order linear equations | 146 |
| 7.33. Second-order equation with the first derivative present | 148 |
| 7.34. Equations of order higher than the second | 149 |
| 7.4. Taylor series method | 149 |
| 7.5. Other procedures | 151 |
| 7.51. Richardson's 'deferred approach to the limit' | 151 |
| 7.52. Iterative processes | 153 |
| 7.53. The Madelung transformation | 154 |
| 7.54. The Riccati transformation | 155 |

| | |
|--|-----|
| 7.6. Two-point boundary conditions | 155 |
| 7.61. Iterative quadrature | 157 |
| 7.62. Linear equations with two-point boundary conditions | 159 |
| 7.63. Factorization method | 161 |
| 7.64. Characteristic value problems | 162 |
| CHAPTER VIII. SIMULTANEOUS LINEAR ALGEBRAIC EQUATIONS | |
| AND MATRICES | 166 |
| 8.1. Direct and indirect methods for simultaneous linear equations | 166 |
| 8.11. Matrices | 168 |
| 8.12. Ill-conditioned equations | 168 |
| 8.13. Normal equations | 171 |
| 8.2. Elimination | 171 |
| 8.21. General elimination process | 173 |
| 8.22. Evaluation of a solution by elimination | 175 |
| 8.23. Alternative arrangement of the elimination process | 178 |
| 8.3. Inverse of a matrix by elimination | 178 |
| 8.4. Choleski's method | 180 |
| 8.41. Inverse of a matrix by Choleski's method | 185 |
| 8.5. Relaxation method | 185 |
| 8.51. Group relaxations | 188 |
| 8.52. Use and limitations of the relaxation method | 189 |
| 8.6. Linear differential equations and linear simultaneous equations | 191 |
| 8.7. Characteristic values and vectors of a matrix | 196 |
| 8.71. Iterative method for evaluation of characteristic values and characteristic vectors of a symmetrical matrix | 199 |
| 8.72. Richardson's purification process for characteristic vectors | 201 |
| 8.73. Relaxation process for characteristic vectors | 207 |
| CHAPTER IX. NON-LINEAR ALGEBRAIC EQUATIONS | |
| 9.1. Solution of algebraic equations | 210 |
| 9.2. Graphical methods | 210 |
| 9.3. Iterative processes | 211 |
| 9.31. Examples of iterative processes | 213 |
| 9.32. Derivation of a second-order process from a first-order process | 216 |
| 9.4. Multiple roots and neighbouring roots | 217 |
| 9.5. Special processes for special types of equations | 218 |
| 9.51. Quadratic equations | 219 |
| 9.52. Cubic and quartic equations | 220 |
| 9.53. Polynomial equations | 221 |
| 9.54. Repeated roots | 222 |
| 9.55. Division of a polynomial by a quadratic | 222 |
| 9.56. Real quadratic factors of a polynomial | 224 |
| 9.57. Second-order process for improving the approximation to a quadratic factor | 226 |

| | |
|---|-----|
| 9.6. Simultaneous non-linear equations | 228 |
| 9.7. Three or more variables | 233 |
| CHAPTER X. FUNCTIONS OF TWO OR MORE VARIABLES. | 235 |
| 10.1. Functions of a complex variable and functions of two variables | 235 |
| 10.11. Numerical calculations with complex numbers | 235 |
| 10.2. Finite differences in two dimensions; square grid | 236 |
| 10.3. The operator $\partial^2/\partial x^2 + \partial^2/\partial y^2$ | 238 |
| 10.31. Special relations when $\partial^2 f/\partial x^2 + \partial^2 f/\partial y^2 = 0$ | 239 |
| 10.4. Finite differences in cylindrical coordinates | 240 |
| 10.5. Partial differential equations | 242 |
| 10.6. Elliptic equations | 244 |
| 10.61. Relaxation process | 245 |
| 10.62. Reducing the mesh size | 249 |
| 10.63. Further notes on the relaxation process | 251 |
| 10.64. Richardson-Liebmann process for Laplace's equation | 253 |
| 10.7. Parabolic equations | 253 |
| 10.71. Replacement of the second-order (space) derivative by a finite difference | 254 |
| 10.72. Replacement of the first-order (time) derivative by a finite difference | 254 |
| 10.73. Replacement of both derivatives by finite differences | 256 |
| 10.74. Note on methods for parabolic equations | 257 |
| 10.8. Hyperbolic equations. Characteristics | 257 |
| 10.81. Finite differences between characteristics | 259 |
| 10.82. Use of given intervals in one independent variable | 260 |
| 10.83. Two simultaneous first-order equations | 261 |
| CHAPTER XI. MISCELLANEOUS PROCESSES | 264 |
| 11.1. Summation of series | 264 |
| 11.11. Euler's transformation for a slowly convergent series of terms of alternate signs | 265 |
| 11.12. Use of the Euler-Maclaurin integration formula in the summation of series | 266 |
| 11.2. Harmonic analysis | 268 |
| 11.3. Recurrence relations for a sequence of functions | 271 |
| 11.4. Smoothing | 272 |
| 11.41. Automatic methods of smoothing | 274 |
| 11.42. Smoothing by use of an auxiliary function | 276 |
| CHAPTER XII. ORGANIZATION OF CALCULATIONS FOR AN AUTOMATIC MACHINE | 279 |
| 12.1. Automatic digital calculating machines | 279 |

| | |
|--|-----|
| 12.2. Preparation of calculations for an automatic digital calculating machine | 283 |
| 12.3. Hand and automatic calculation | 284 |
| EXAMPLES | 287 |
| BIBLIOGRAPHY | 293 |
| INDEX | 299 |

I

INTRODUCTION

1.1. What numerical analysis is about

THE subject of numerical analysis is concerned with the science and art of numerical calculation, and particularly with *processes* for getting certain kinds of numerical results from certain kinds of data. The following are some simple typical problems for which we may require processes for obtaining numerical solutions:

(i) Tabulate $(\sinh x - x)/x^3$ to five decimals for $x = 0(0.1)3$.†

(ii) Given such a table,

(a) find, as accurately as possible, the value of x for which

$$(\sinh x - x)/x^3 = 0.2;$$

(b) construct a table at intervals of 0.02.

(iii) What values of x, y, z satisfy the equations

$$xyz = 6, \quad x^2 - y^2 + z^2 = 6, \quad x + 2y + 3z = 10?$$

(iv) Tabulate $\int_0^\infty e^{-(x-w)^2} dw$ for $x = -2(0.1)2$.

(v) For what values of λ has the equation

$$y'' + (\lambda - e^{-x})y = 0$$

got a solution for which $y \rightarrow 0$ as $x \rightarrow \pm\infty$ and for which

$$\int_{-\infty}^{\infty} y^2 dx = 1,$$

and what are the corresponding solutions?

Although from the point of view of numerical analysis the end to be attained is always a numerical result or set of results, the subject is not concerned with the *results*, that is to say *answers* to specific problems themselves, but with the *processes* by which those results can be obtained. And although the end is a numerical result, algebra and analysis are involved in the development of these processes. In so far as these processes, and the arguments by which they are derived, are general and independent of the particular values of the numbers to which they may be

† This is a standard notation for 'from $x = 0$ to 3 inclusive at intervals of 0.1 in x '; see § 2.3.

applied, the subject may properly be regarded as a branch of mathematics.† But the algebra and analysis must be aimed at providing or establishing *practical methods of obtaining numerical results*; otherwise it may be elegant mathematics, but is not a contribution to numerical analysis.

This emphasis on *practicable numerical processes* requires a considerable change in attitude from that of ordinary algebra and analysis, to which the idea is quite foreign. Algebraical or analytical results which are formally complete answers may be almost or quite useless for numerical purposes. Consider, for example, the solution of a system of simultaneous linear algebraic equations. Any textbook of algebra shows how this can be expressed in terms of ratios of determinants, and this result is often presented in a form which seems to imply that there is nothing more to be said on the subject. But direct evaluation of the solution in this form is certainly not the practical answer to the problem of finding a *numerical* solution of a set of simultaneous equations.

As another example, consider the solution of

$$\frac{dy}{dx} = 1 - 2xy, \quad y = 0 \text{ at } x = 0. \quad (1.1)$$

The standard textbook treatment gives

$$y = e^{-x^2} \int_0^x e^{x^2} dx, \quad (1.2)$$

and regards this as a complete answer. And so, for numerical purposes, it is, *provided* one has a table of $\int_0^x e^{x^2} dx$. But in order to obtain such a table it is much easier to reverse the process and solve the differential equation (1.1) by numerical methods, and then to evaluate the integral by using (1.2) in the form

$$\int_0^x e^{x^2} dx = e^{x^2} y,$$

than to evaluate $\int e^{x^2} dx$ numerically directly. Again the formal textbook answer is of no practical use if numerical results are wanted in the end. As in these two examples, practical numerical considerations which are irrelevant to formal mathematics may require alternative methods

† A. N. Whitehead has written (*Introduction to Mathematics*, p. 15): 'Mathematics as a science commenced when first someone, probably a Greek, proved propositions about *any* things or *some* things, without specification of definite particular things.' The *methods* of numerical analysis, as distinct from the details of their application in particular cases, have that degree of generality which entitles them to be considered part of mathematics.

for treating problems for which complete formal solutions may already be known.

Another matter in which there is much greater emphasis in numerical analysis than in formal analysis is the checking of numerical work. Numerical results which are not reliable are of little or no value, and for this reason any process for obtaining them should include checking procedures for confirming that the alleged results are free from mistakes. This is considered further in § 1.3.

As already mentioned, numerical analysis is concerned with *processes*. It is an active subject, one in which things *happen* in the course of carrying out numerical processes, and it cannot be learnt properly simply by reading about it, by following examples already worked, or even by watching examples being worked, any more than one can learn golf, tennis, or violin-playing by watching others play, without ever handling a club, racket, or violin. There is a great deal of difference between only *thinking* about processes for carrying out numerical calculations and actually carrying them out with numbers in the place of the algebraical symbols of a general treatment, and the student who wishes to get a feeling for the subject *must* work examples of the processes for himself. This is an essential part both of study and research in the subject. Also, probably, he must make his own mistakes and spend time finding them and correcting them and their consequences before he really appreciates the importance of adequate checking.

The processes of numerical analysis are necessarily *finite* processes. Ideas such as limiting processes, Dedekind sections, formal convergence, scarcely play any part in the numerical processes themselves, though they may be involved in the analytical arguments by which the numerical process is established. Related to this is the approximate nature of much of numerical analysis. In most applications of numerical analysis, almost no problems have answers which are rational numbers. But our system of representation of numbers is not suitable for numerical operations on irrational numbers, so that in most cases we have to be satisfied with approximations. And even when the answers are rational numbers, we shall often be content with decimal approximations to these rational numbers, if indeed we would not prefer them.

It will be as well to end this section by explaining what numerical analysis is *not*.

First, it is *not* necessarily concerned with the analysis of numbers obtained by observation in the course of some branch of experimental science; secondly, it is *not* closely related to statistics. Certainly numerical

analysis may be involved in the analysis of observational material, whether statistical or obtained by measurement, and the analysis of observations consisting of measurements may involve consideration of the statistics of errors of the measurements. But the subject itself is distinct from these two particular applications of it, just as it is distinct from its particular applications, for example, to the evaluation of supersonic fluid flow or to the structures of atoms or stars.

1.2. The main types of problems in numerical analysis

The main kinds of operations which have to be carried out in the course of a numerical calculation, and for which numerical processes are required, are the following:

- (a) Evaluation of formulae.
- (b) Solution of non-linear equations in one unknown.
- (c) Solution of systems of linear simultaneous equations.
- (d) Inversion of matrices.
- (e) Determination of characteristic values and characteristic vectors of matrices.
- (f) Solution of systems of non-linear simultaneous equations.
- (g) Tabulation of standard functions.
- (h) Interpolation and subtabulation.
- (i) Integration and differentiation of a given function.
- (j) Smoothing.
- (k) Integration of ordinary differential equations.
- (l) Integration of partial differential equations.
- (m) Solution of integral equations.
- (n) Harmonic analysis.
- (o) Frequency analysis (periodogram analysis).

Of these, (j), (n), and (o) are often concerned with analysis of *observed data*, which is not primarily the subject of numerical analysis as pointed out at the end of the previous section.

A single calculation may involve a number of these operations. For example, evaluation of the solution of an ordinary differential equation may well involve any one or more of (a), (b), or (h) as well as the integration process (k) itself.

The subjects in this list have been arranged more or less in order from less to more 'advanced'; they will not, however, be taken in this order, since some of the ideas required in treating later items of this list are also valuable in the earlier ones.

1.3. Errors, mistakes, and checking

There are three reasons for which the results of a numerical calculation may differ from the exact answer to the mathematical question concerned:

- (i) One (or more) of the formulae which are evaluated in the course of the work is derived by cutting off an infinite series after a finite number of terms; the errors introduced in this way are called 'truncation errors';
- (ii) Only the more significant decimal digits of a number are retained, the less significant beyond a certain point being rejected: this process is called 'rounding off' and the errors introduced in this way are called 'rounding errors' or 'rounding-off errors';
- (iii) Mistakes are made in carrying out the sequence of operations required to obtain the results sought.

The distinction made here between an 'error' and a 'mistake' is this. A 'mistake' is due to fallibility, either human on the part of the individual carrying out the calculation, or technical on the part of the mechanical or electrical aids used in the course of it, and is in principle avoidable. 'Errors', in some degree, are unavoidable, except in some cases of calculations concerned entirely with integers or rational numbers; such calculations may occur, for example, in connexion with number theory, but are otherwise exceptional. 'Truncation errors' are unavoidable in any process which takes the place in numerical work of a limiting process of analysis; integration and differentiation are two important examples. 'Rounding errors' are inevitable in division when the answer is a non-terminating decimal, and in the use of values of functions other than polynomials with rational coefficients, and are often incurred in multiplication also, since although it is possible to retain the $(m+n)$ digits of the product of two numbers, one of m and the other of n digits, it is only exceptionally that all these digits are wanted; if one or both of the numbers being multiplied is subject to rounding error, some of the less significant digits in the product will be valueless anyway.

It is necessary to check that the final results of a calculation are not vitiated either by errors or by mistakes, and in a substantial calculation it will usually be advisable to include a number of checks of intermediate results as well. It is often possible to estimate the magnitude of truncation errors and so ensure that they are kept below a specified tolerance depending on the calculation and the accuracy required in the final results. Rounding errors can often be rendered innocuous by carrying one or two,

or sometimes more, extra figures, known as 'guarding figures', in intermediate stages of the calculation; for example in calculating a compound interest table of

$$f(p) = (1.0325)^p$$

for $p = 0(1)100$, to five decimals, by repeated use of the recurrence relation

$$f(p+1) = 1.0325f(p),$$

rounding errors greater than 6 in the sixth decimal can be avoided by keeping eight decimals in the intermediate values of $f(p)$. A full analysis of the effect of rounding errors in any but a simple calculation may be fairly elaborate.

Intermediate and final results of a calculation will usually be influenced by rounding errors at previous stages of the work, and in some cases the accumulated effects of rounding errors will result in checks not being satisfied exactly. Let y be the correct value of a quantity and y^* the calculated value of it. Then there may be a range of values of $y - y^*$ which can be accepted as being results of rounding (and possibly truncation) errors and not as indicating mistakes. The term 'tolerance' (in the sense in which it is used in machining work in engineering) will be used for this acceptable range of $y - y^*$. For example, if a check consists of the equality of two numbers calculated by different processes, and the tolerance of each is ± 2 in the last digit, a difference of 3 between them in this digit is within the tolerance on this difference, and can be passed.

Anyone intending to undertake a serious piece of calculation should realize that adequate checking against *mistakes* is an essential part of any satisfactory numerical process. No one, and no machine, is infallible, and it may fairly be said that the ideal to aim at is not to avoid mistakes entirely, but to find all mistakes that *are* made, and so free the work from any *unidentified* mistakes. This of course is an ideal. It does not seem possible to eliminate mistakes with absolute certainty; it is always possible that a mistake might be made in the check itself in such a way as to cancel the effect of an error it was devised to find. But with properly designed checking procedures and care in working, the probability of this should be negligibly small.

Provision of adequate checks is not, however, to be regarded as an excuse for mistakes or a justification of carelessness in carrying out the details of numerical work. Location and diagnosis of a mistake, and correction of the mistake itself and of subsequent calculations vitiated by it, is often a time-consuming job, and a tiresome one at that; and moreover, if mistakes are too frequent, the probability of a mistake in a check masking a mistake which the check should detect may become

appreciable. Numerical work should always be done with care to avoid mistakes, and checks regarded as insurance against the occasional mistakes which may occur even in careful work.

Many calculations consist of the same group of arithmetical operations applied repeatedly to different data. For example, if it were required to evaluate the function y defined by

$$y = \frac{1}{2}x^2 + \frac{3}{5!}x^5 + \frac{3 \cdot 6}{8!}x^8 + \frac{3 \cdot 6 \cdot 9}{11!}x^{11} + \dots \quad (1.3)$$

for a set of values of x , say $x = -3 \cdot 0(0 \cdot 1)3 \cdot 0$, by evaluating and summing the separate terms of the series, the process of calculating y is the same for each value of x (except that for the smaller values of x more terms of the series are negligible and do not have to be evaluated explicitly). Such a systematic set of calculations is easier to check than one in which no step is similar to any other. A *single* value of the function y would be difficult to check adequately; a systematic set of values can be checked comparatively easily. In this case, for instance, a check might be based on the fact that y defined by (1.3) is a solution of the differential equation $y'' = 1 + xy$ (for an example, see § 3.3), but use of such a check depends on the behaviour of y as a function of x , and is not applicable to a single isolated value of y .

Mistakes in such a calculation are of two kinds, systematic (that is, the same mistake is made at the same point in each repetition of the sequence of arithmetical operations) and random. These can be illustrated from one method of evaluating the above series (1.3). Suppose the $(n+1)$ th term is evaluated by multiplying the n th by $x^3/(3n+1)(3n+2)$; then

$$(\text{third term}) = (x^3/56) \times (\text{second term}),$$

and the denominator here might be taken as 54 instead of 56 throughout the whole calculation for all values of x ; this would be a systematic mistake. On the other hand, one too many or one too few zeros between the decimal point and the first significant figure might be taken in a single one of the terms of the series for a single value of x ; this would be a random mistake.

It is recognized by those with extensive experience of numerical work that there are two kinds of random mistake which are particularly easy to make. One is an interchange of adjacent digits; for example, 28575 may be read or recorded as 25875. The other is repeating the wrong digit in a number in which two adjacent digits are the same; for example 36609 may be read or written as 33609 or 36009. The error introduced by a mistake of the first of these kinds is always a multiple of 9 in terms of the

less significant of the two interchanged digits as unit; this may often help in locating and identifying a mistake of this kind. These are not, of course, the only kinds of mistakes that can be made: but if a check indicates the presence of a random mistake, knowledge that it is likely to be of one of these kinds may assist in diagnosing it.

In the calculation of the function y defined by (1.3) by evaluation of the series, the calculations for different values of x are independent, so that a mistake in the calculation for one value of x does not affect those for later values of x .

But in many calculations, such as a calculation of this same function y by numerical integration of the differential equation $y'' = 1 + xy$ satisfied by it, a mistake at one stage vitiates all subsequent work. In such a case, it is important to have a *current* check on the work as it progresses rather than only an overall check carried out when the calculation is completed, otherwise the amount of work that has to be repeated if a mistake is made may become very considerable. All the time spent on work subsequently found to be vitiated by a mistake is just wasted, and a few experiences of this kind may be found severely discouraging, although really the moral should be simply that an adequate current check is needed.

One kind of 'check' is so inadequate as to be almost worthless, namely, repetition of a calculation by the same individual that did it originally. It is much too easy to make the same mistake twice; and indeed it may be that having made a mistake once, one is conditioned to make it again on repeating the work. An *independent* repetition of the work by a second individual is better than no check, but should not be regarded as adequate. The only really satisfactory check is one which obtains or verifies a result by a *different sequence of arithmetical operations, or a sequence involving different numbers, from that by which it was obtained*. For example, values of $\cosh x$ and $\sinh x$ interpolated from tables may be checked by use of the identity $\cosh^2 x - \sinh^2 x = 1$ (this does not check that they are not both interpolated for the wrong value of x , but this can probably be checked in some other way, depending on the rest of the calculation for which values of $\cosh x$ and $\sinh x$ are wanted); and the values of y calculated from the series (1.3) can be checked by use of the differential equation satisfied by y .

1.4. Arrangement of work

In most numerical work, a working sheet will be used for recording data and intermediate results of the calculation. A clear and orderly arrange-

ment of this working sheet is a great help both in avoiding mistakes and in locating and correcting any that do happen to be made. Numerical work should not be done on odd scraps of rough paper, but laid out systematically and in such a way as to show how the intermediate and final results were obtained; and the numbers entered on the work sheet should be written neatly and legibly. Use of ruled paper is a help in keeping the layout of the work neat and clear. It is advisable to use loose sheets rather than a book since it is rather easy to make mistakes in copying from one page to another of a book; with loose sheets the number to be copied from one sheet, and the place to which it is to be copied on another, can more easily be brought close together, and the copy made and checked more easily.

For work of any permanent value, it is advisable to record on the working sheet enough explanation of the different entries, and how they were obtained, for the working to be followed after the lapse of a period of years.

1.5. Accuracy and precision

In contexts in which numerical work is carried out in connexion with scientific and technical problems, we are often concerned with the numerical solution of one or a set of algebraic, differential, or integral equations. Then it may be convenient to distinguish between the accuracy to which the equations, or data used in obtaining a solution of them, represent the real situation to which they refer, and the accuracy to which the results of the numerical work represent the solution of these equations with these data, supposed exact. The latter is sometimes distinguished by being called the 'precision' or 'nominal accuracy' of the numerical work.

Calculations are often carried out deliberately to a nominal accuracy known or expected to be higher than the accuracy of the approximations made in deriving the equations, or higher than that of the data used in their solution. There are several reasons why this may be done. We may be interested in the differences between the results of observation and of calculation, whether for the purpose of assessing the accuracy to which the equations do give an account of the observations, or in order to analyse these differences so as to derive more accurate equations or data to use in them. Then we want to be sure that the differences between the results of observation and of calculation are significant, and are not merely consequences of the limited nominal accuracy of the calculations.

Or we may want to determine the difference between two solutions of

the equations with different values of some parameters, and to obtain this difference we may have to calculate the separate solutions to a nominal accuracy higher than that of the data. Or the results may be only intermediate results on which some extensive interpolation, perhaps in two or three variables, is going to be carried out. Both for the interpolation process and for checking purposes, it is then desirable that these intermediate results should be smooth and of a nominal accuracy higher than required in the final results.

In a hand calculation, however, greater nominal accuracy means more work, more writing in recording intermediate results, more possibilities of mistakes, and a longer time for the calculation. It is advisable, therefore, to watch lest the calculation is being carried to an unnecessarily high nominal accuracy. In this connexion a warning may be given concerning the use of desk machines. Since with a desk machine it is possible to work to eight or ten figures, there is a tendency to get into a habit of working with eight or ten figures when four or five would be adequate. This is bad practice, and a habit which the serious student of numerical work should avoid for his own sake.

II

THE TOOLS OF NUMERICAL WORK AND HOW TO USE THEM

2.1. The main tools of numerical work

FOR carrying out the numerical details of a calculation there are four main kinds of tools:

- | | |
|--------------------|------------------|
| (a) Desk machines. | (c) Slide rule. |
| (b) Tables. | (d) Graph paper. |

Of these the first and second are much the most important.

2.2. Desk machines

A desk calculating machine is the most important single tool for numerical work and anyone intending to study numerical analysis seriously should become familiar enough with the main kinds to use them with facility, without more deliberate thought for the details of operating the machine than a good typist gives to the operation of individual keys of the typewriter.

There are several kinds of desk machine, some being primarily adding machines whereas others have facilities for multiplication; the former are sometimes called 'adding machines' as distinct from 'calculating machines' to emphasize this feature. The latter are the more important and will be considered first; adding machines are considered in §2.26. Of the calculating machines some are considerably different in appearance and operation from others, but all are broadly similar in general principle. All have four main components:

- (i) A setting mechanism by which a number can be set on the machine.
- (ii) A register in which results of additions, subtractions, and multiplications are accumulated; this will be called the 'accumulator'; other names for it are 'result register' and 'product register'.
- (iii) A counting register, sometimes called 'multiplier register', on which a count is kept of the number of additions or subtractions made.
- (iv) An operating handle (in hand machines) or key-operated switch (in electrically-driven machines).

The setting mechanism and registers have means for setting them to zero; this is called 'clearing'.

Three kinds of desk calculating machines are illustrated in Figs. 1, 2, and 3. Three different kinds of setting mechanisms are represented in these three machines, and this is the main reason for the difference in appearance between them.

On the Brunsviga (Fig. 1), the setting mechanism consists of a series of levers, one for each digital position, each lever having ten positions corresponding to the decimal digits 0–9. On the Marchant (Fig. 2) the setting mechanism consists of a keyboard on which there is a set of nine keys, corresponding to the digits 1–9, in a column in each digital position; a number is set by pressing the appropriate key in each column. On the Facit (Fig. 3) there is a keyboard of only ten setting keys, corresponding to the digits 0–9; a number is set by pressing these keys in an order corresponding to the order of the digits in the number, beginning with the most significant.

The accumulator can be traversed relative to the setting mechanism, so that the least significant digital position of the adding mechanism corresponds to different digital positions of the accumulator. Shifting the accumulator one place to the *right* corresponds to multiplication by 10.

On all these machines multiplication is carried out by repeated addition and shifting. Machines which carry out multiplication directly, by use of a built-in multiplication table, have been constructed, but some machines carrying out multiplication by repeated addition are now so fast that, for work in which numbers are supplied manually to the machine, there is little purpose in making machines using direct multiplication.

On hand-operated machines, addition is carried out by rotating the handle through one turn in one direction (clockwise, looking along the handle towards the body of the machine) and subtraction by rotating it through one turn in the other direction. Most machines have a lock on the handle so that once a turn has been started it must be completed, and often have mechanical interlocks to prevent incorrect operation. On an electrically-driven machine the rotation is supplied by an electric motor instead of directly by the operator's hand, and the motor is controlled by a set of key-operated switches.

This is not the place for an account either of the internal mechanism or of the details of operation of different machines; the operating procedure is given in booklets supplied with the machine or obtainable from the makers or agents, but probably can best be acquired from personal demonstration by someone already familiar with the machine. The

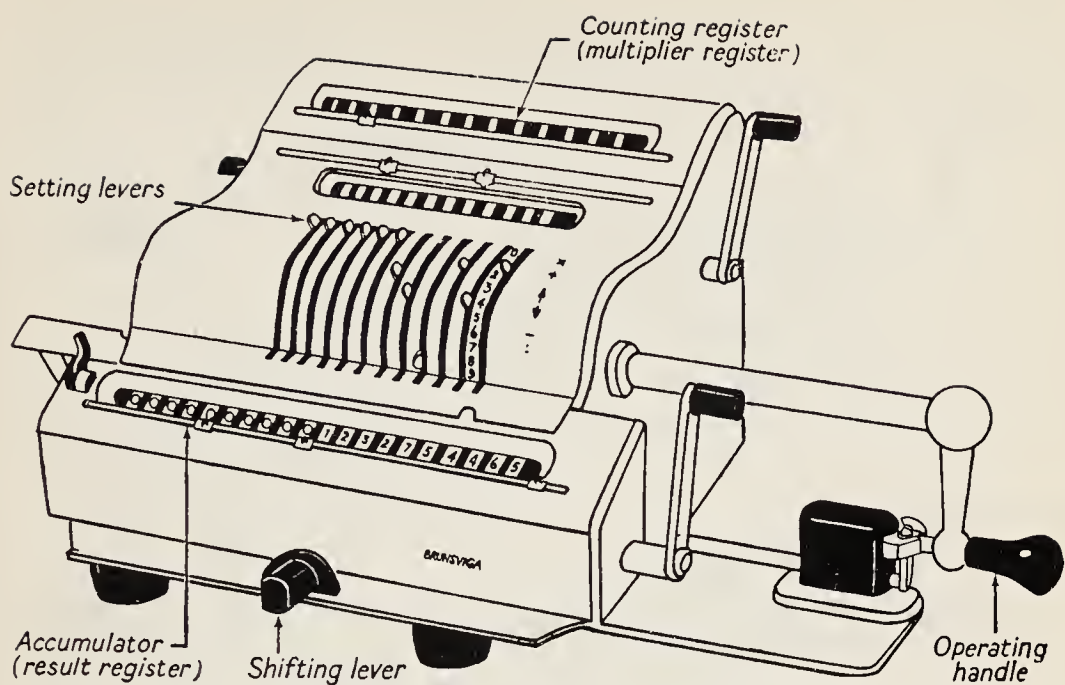


FIG. 1. Brunsviga (hand-operated).

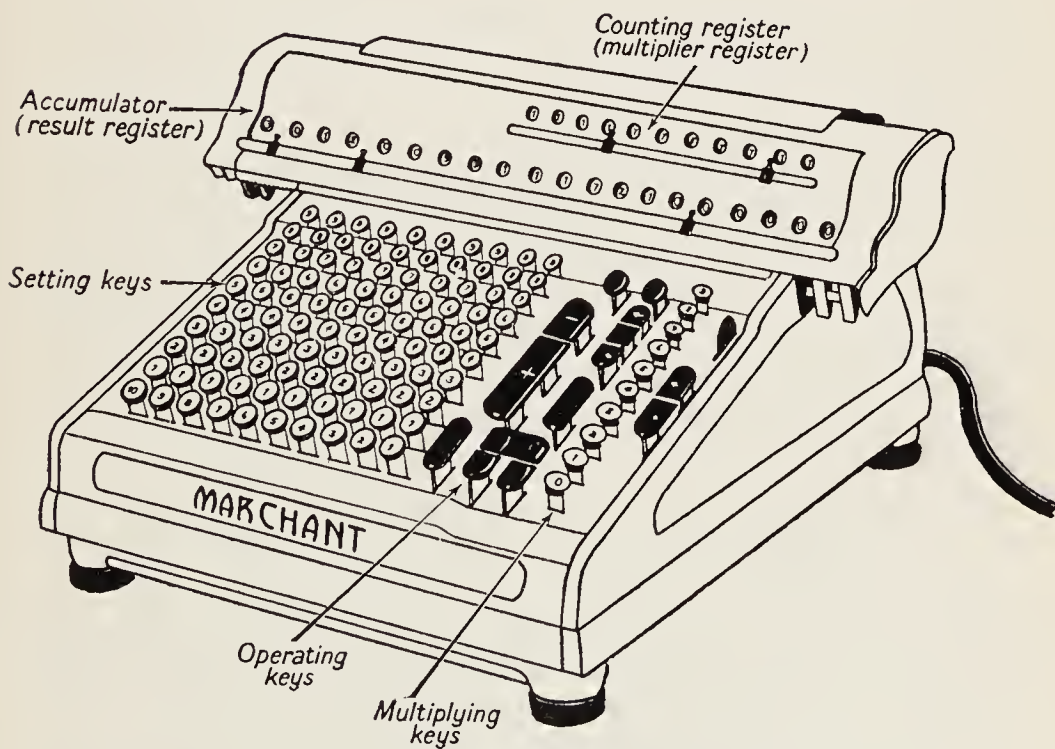


FIG. 2. Marchant (electrically operated).

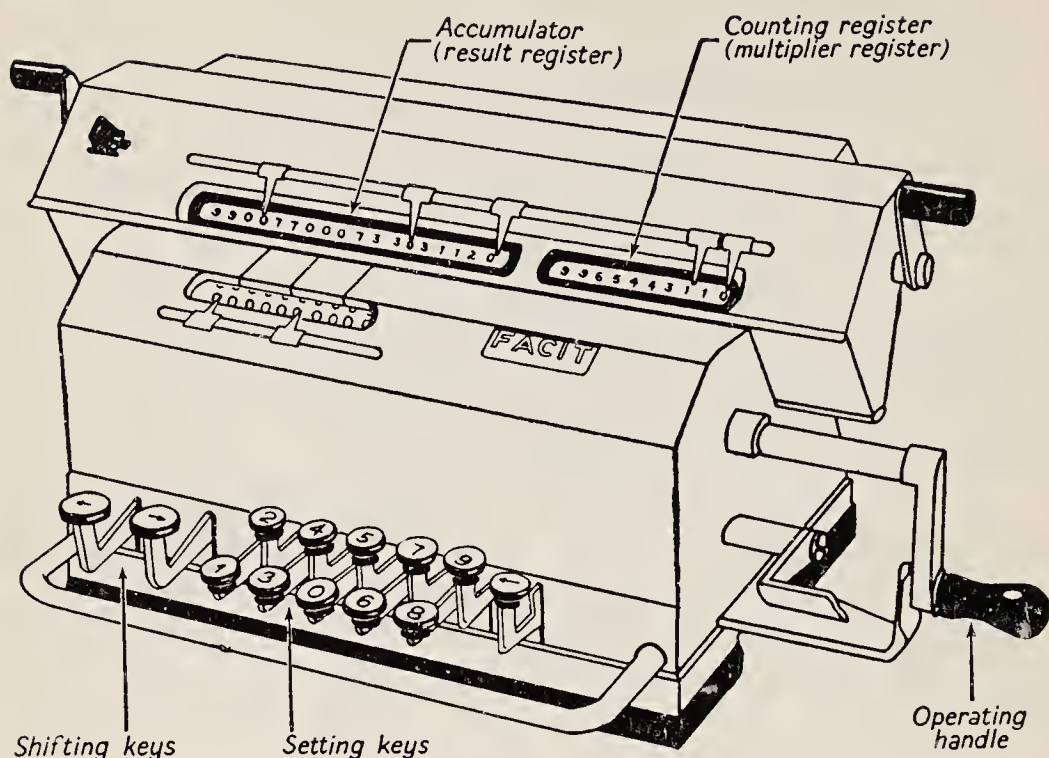


FIG. 3. Facit (hand-operated model).

following sections deal with some general points of procedure applicable to most machines.

2.21. Addition and subtraction

Addition is carried out by setting the addend on the setting levers or keys and turning the handle once positively, or on an electric machine by pressing the $+$ key or by multiplying by 1 with the shift-control set to 'non-shift'; on some machines the latter procedure is necessary when it is required to hold the number set, since this is cleared after addition when the $+$ key is used. The number set up is then added to the content of the accumulator. The position of the decimal point needs watching if the number of digits after the decimal point is different in the addend and in the content of the accumulator. Decimal-point markers are furnished on all machines (their form is different on different machines); in single arithmetical operations it is often unnecessary to use them, but they are very useful in helping to keep the position of the decimal point correct in carrying out sequences of operations on the machine without writing down intermediate results, as is sometimes possible.

Subtraction is carried out similarly to addition, except that the handle is turned in the opposite direction. The result of subtracting a greater

number from a smaller is as follows. Let \bar{a} be written for a contribution ($-a$) in any digital position, so that, for example, the number 90 can be written $1\bar{1}0$ and the number 88 as $1\bar{1}\bar{2}$. Then the negative number -23 (for example) which is

$$-23 = -1,000 + 977 = -10,000 + 9,977 = -1,000,000 + 999,977$$

can be written

$$-23 = \bar{1}977 = \bar{1}9977 = \bar{1}999977, \quad \text{etc.} \quad (2.1)$$

The number 999...99977, to the full capacity of the accumulator of the machine, is called the 'complement' of 23, or the 'complementary form' of the number -23 ; it can be regarded as a representation of the number -23 in the form (2.1), with the digit $\bar{1}$ to the left of the most significant digital position of the accumulator. In a number in complementary form, the digits to the right of the row of 9's are the significant figures.

Negative results appear in such a complementary form, and, in some machines, a carry-over from the most significant digital position of the accumulator is indicated by the ringing of a bell.

Recording of negative numbers will usually be in terms of sign and *modulus*, not in their complementary form. The translation from the complementary form to the modulus can be done in two ways:

- (i) Translate mentally by subtracting each digit of the complement *except the last* from 9 and subtracting the last from 10. Set the result on the setting levers or keys, add into the accumulator, and *verify that the content of the accumulator is now zero*. This checks the translation and should *always* be done before the result is recorded. If the number in complementary form is wanted in the accumulator for further numbers to be added to it, it can be recovered by subtracting the number on the setting levers.
- (ii) Transfer the number in complementary form from the accumulator to the setting levers or keys (see § 2.22), and subtract from zero. This will give some spurious 9's on the extreme left of the accumulator, but it will be easy to distinguish these from the significant figures of the result.

2.22. Transfer from accumulator to setting keys or levers

In some calculations it is necessary to transfer to the setting levers or keys a number formed in the accumulator as the result of previous calculations. For example, in the calculation of a continued product, an intermediate product formed in the accumulator has to be transferred to the setting levers or keys to be ready for multiplication by the next factor; and as already mentioned in the previous section, such a transfer is a step

in one method of obtaining the modulus of a negative number expressed in complementary form.

Some machines are provided with facilities for direct transfer from accumulator to setting levers or keys. In using one that is not, the following procedure should be followed.

Copy on to the setting keys or levers the number to be transferred, *subtract it from the content of the accumulator, and verify that the result is zero*. This checks that the number has been copied correctly on to the setting mechanism, and this check should *always* be used.

On an electric machine, the subtraction must be done in such a way as *not* to clear the keyboard after subtraction.

2.23. Multiplication

Multiplication is carried out by repeated addition in each digital position of the multiplier, the accumulator being traversed one place right or left between successive digits of the multiplier. In most cases it is best to carry out multiplication starting with the *most* significant digit of the multiplier, as then the order of the digits is the natural one, in which it is easy to remember the multiplier while the multiplication is being carried out.

In a few machines, mainly older models, in which the mechanism for carrying-over in addition does not extend to the full capacity of the accumulator, this procedure will occasionally lead to incorrect results. The best way to test whether a machine has this objectionable feature is to subtract 1 from 0 with the accumulator in the extreme left position, and see if the carry-over produces 9's right to the extreme left-hand digital positions of the accumulator. If not, the best way of avoiding trouble is not to use such a machine; but if no other is available the possibility of incorrect results of this cause must be kept in mind. In multiplication they can be avoided by starting from the least significant digit of the multiplier, but this is inconvenient as it means taking the digits in the opposite order to that in which they will naturally be remembered.

Appreciable time can be saved in multiplication on a hand machine by a procedure known as 'short-cutting'. If, as in § 2.21, a bar over a digit is used to represent a negative digit *in that digital position only*, we have, for example:

$$\left. \begin{array}{ll} 183 = 2\bar{2}3 & (\text{l.h. 12, r.h. 7}) \\ 2879 = 3\bar{1}2\bar{1} & (\text{l.h. 26, r.h. 7}) \\ 369175 = 43\bar{1}2\bar{2}\bar{5} & (\text{l.h. 31, r.h. 17}) \end{array} \right\}. \quad (2.2)$$

Multiplication by one of these numbers can be carried out by using as multiplier the number in the form given on the right-hand side of the equalities in (2.2), and using both positive and negative directions of turning the handle; an appreciable number of turns may be saved in this way; this is the process of 'short-cutting'. The numbers of turns taken to carry out a multiplication by each of the numbers in the example (2.2), in its form on the left-hand side and in its form on the right-hand side of the equality sign, are shown in brackets. 'Short-cutting' should be used on digits over 5, and on a 5 if flanked on either side by a digit over 5; for users of hand machines, it should become the natural way of carrying out multiplications; it needs a little practice at first to become proficient and reliable, but ease in using it is certainly worth attaining.

In a few old models of machines, which have not carry-over (sometimes called 'tens-transmission') mechanism in the multiplier register, only the moduli of the individual digits are indicated (in some machines the negative digits are indicated in red). Such machines should be avoided, or, if they have to be used, short-cutting must be used with discretion and particular attention should be paid to checking.

Some electric machines are fitted with means by which multiplication by any digit of the multiplier and the succeeding shift of the accumulator can be carried out by pressing one of a set of ten keys; these machines are so fast that short-cutting is unnecessary. On others the complete multiplier can be set and transferred to a register, then the multiplicand set and the multiplication carried out automatically; in these machines the operator is not concerned at all with the process of multiplication by individual digits.

2.24. Division

Division can be carried out in three ways:

First, by multiplication by the reciprocal of the divisor. This is particularly useful when the result of the division is required to be in the accumulator, either in order to have further numbers added to it or for transfer to the setting levers or keys. In the other methods of division, the quotient appears in the multiplier register, and no machine has transfer facilities from there to the setting levers or keys; this transfer had to be done by hand and there is no adequate means of checking it, whereas if a result is in the accumulator its transfer to the setting levers or keys can be made mechanically or checked (see § 2.22).

Secondly, by successive subtraction, starting from the *most* significant digit. In this process, the dividend, if not already in the accumulator as

a result of previous operations, is set and added into the accumulator which has previously been cleared. *The multiplier register must then be cleared* (this is automatic in the case of automatic division on some electrical machines); this is a step which is rather easily overlooked. The divisor is then set and subtracted in the most significant position until the remainder is less than the divisor; the accumulator is then shifted one place left, and the subtraction followed by a shift is repeated. The result appears in the multiplier register. In order to make full use of the capacity of this register, the divisor should normally be set in such a position on the setting levers or keys that the quotient has a non-zero digit in the extreme left digital position of the multiplier register.

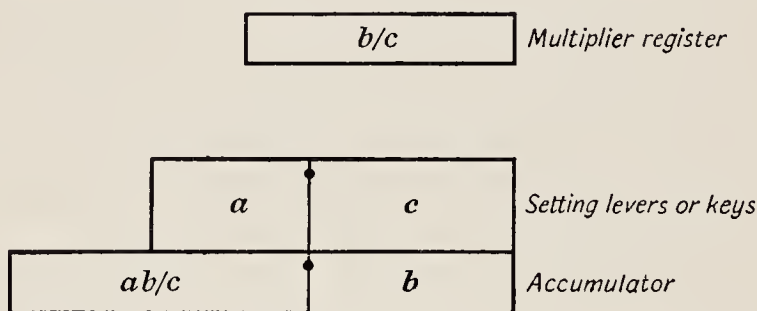


FIG. 4. (Dots • show decimal-point markers used as separators.)

Thirdly, by successive addition (sometimes called division by 'building up'). In this process the accumulator is cleared, the divisor set on the setting levers or keys, and *multiplied* by such a number x that the result in the accumulator is the dividend. This multiplication is done by a process which is essentially one of trial, but can be quite fast, and in which short-cutting can be used to some extent. It is useful when the same divisor is used with a number of dividends, as then this divisor can be set up once for all and need not be disturbed to set the new dividend. It is sometimes also useful for forming ab/c in one operation, if the number of digits involved is not too large. If a is set at one side of the setting mechanism and c at the other, and b/c is found by building up c to b , then the handle has been turned a number of times corresponding to (b/c) , and a has been multiplied by that number. This arrangement is shown diagrammatically in Fig. 4. Another application of the process of division by building up is in the calculation of $|a/b|$, where a is a negative number standing in the accumulator (in complementary form) as the result of a previous calculation. If $|b|$ is set, and the content of the accumulator built up to zero, the multiplier register will read $|a/b|$.

Most electric machines are provided with facilities for automatic division; the dividend is placed in the accumulator, either by adding it in after clearing the accumulator, or by forming it there as a result of previous calculations. The divisor is then set and the 'automatic division' key pressed; the division then proceeds without further manipulation on the part of the operator.

2.25. Other calculations

With a machine of sufficient capacity, and numbers of a few digits, it is possible to do two calculations simultaneously, one with numbers set on the extreme left and the other with numbers set on the extreme right of the setting levers or keyboard; an example has already been given in the calculation of ab/c in one operation. The following are two other examples:

- (i) $\sum_n a_n$ and $\sum_n a_n b_n$ *simultaneously*. Set 1 on the extreme left and the numbers b_n successively on the extreme right; for each b_n multiply by the corresponding a_n . Then in the accumulator $\sum_n a_n$ is formed on the left and $\sum_n a_n b_n$ on the right.
- (ii) $\sum_n a_n^2$ and $\sum_n a_n b_n$ *simultaneously*. Set a_n on the extreme left and b_n on the extreme right, multiply by a_n , and repeat for each value of n . Then in the accumulator $\sum_n a_n^2$ is formed on the left and $\sum_n a_n b_n$ on the right. If the multiplier register is *not* cleared between each multiplication, $\sum_n a_n$ is accumulated there, but this is hardly satisfactory, as it is then impossible to check after each multiplication that the right multiplier has been used. An overall check can be provided by setting the pairs of numbers a_n, b_n in succession and multiplying each pair by the corresponding b_n . This would give $\sum_n a_n b_n$ and $\sum_n b_n^2$; the latter is likely to be wanted in contexts in which $\sum_n a_n^2$ and $\sum_n a_n b_n$ are wanted, and the agreement of the two values of $\sum_n a_n b_n$ would check that the right multiplier values had been used in each calculation; it does not check the setting of the values of a_n in the first, or of b_n in the second, of the calculations.

2.26. Adding machines

In adding machines the position of the accumulator relative to the keyboard is fixed, and there is no multiplier register. Most of them have

keyboard setting, and many have electrical operation controlled through a set of keys.

The most useful of these machines are those which make a printed record of each number added into the accumulator. There are two operations by which a total standing in the accumulator can be printed. If a key marked 'total' is operated, the total is printed and the accumulator is *cleared*; if a key marked 'sub-total' is operated the total is printed and retained in the accumulator. A particular application of the latter

operation is in the evaluation of an integral $\int_a^x f(w) dw$ as a function of its upper limit x , by successive addition of contributions from successive intervals of x . After each contribution is added, a sub-total is taken, then the next contribution is set and added. The printed record consists of a sequence of entries, alternately contributions to the integral and values of the integral itself. The contributions actually used by the machine can then be checked against the values which should have been set.

It should be a convention in using a machine of this kind that it is left with the accumulator clear; but in case this has not been done, it is advisable, before using it, always to ensure that the accumulator is clear by taking a total.

2.3. Mathematical tables

Mathematical tables† form a very important aid to numerical work. Many calculations involve the use of values of standard functions such as exponentials, logarithms, circular functions, Bessel functions, the gamma function, and though it would be possible to calculate the required function values from scratch as they were wanted, this would usually lengthen the calculation so much as to make it impracticable. In fact, if tables of these functions did not already exist, it would often be worth constructing them as a first step in the calculations for which values of these functions are wanted.

The most important tables are the following:

Comrie and Milne-Thomson's Standard 4-Figure Tables; *Chambers's 6-Figure Mathematical Tables* (2 vols. 1948-9) and *Chambers's Shorter 6-Figure Tables* (1950), edited by Comrie; *Barlow's Tables of Squares, Cubes, Reciprocals, etc.*, edited by Comrie; *Interpolation and Allied Tables* (H.M. Stationery Office, 1956).

† On this general topic, see L. Fox, *The Use and Construction of Mathematical Tables* (H.M.S.O., 1956).

The first two of these include tables of circular functions for argument in radians, and also tables of inverse trigonometric and hyperbolic functions; of the two volumes of Chambers's 6-figure tables, the second with so-called 'natural' values is much the more useful for work with machines. *Interpolation and Allied Tables* contains a great deal of information on formulae and methods for interpolation and other numerical processes and is a very useful and inexpensive booklet.

For functions other than the elementary functions, the following are useful:

Dale, *5-Figure Tables of Mathematical Functions*; Jahnke-Emde, *Tables of Functions with Formulae and Graphs*; British Association *Mathematical Tables*, Vols. VI, X and Part-vols. A, B.

The amount of tabular material available in various volumes of tables and scattered among various journals is very considerable. The nature and location of most of this material published up to the end of 1944 has been classified and tabulated in an *Index of Mathematical Tables*† which is a most valuable volume and should be known to all undertaking any extensive numerical work, or even small calculations involving functions other than the elementary functions, since if a function has been tabulated, knowledge of this fact and an adequate reference will usually avoid duplicating the calculation of it.

An important source of information, particularly regarding recent or current work on tabulation of functions, is the journal *Mathematical Tables and Aids to Computation* (generally referred to as *M.T.A.C.*).

In describing a table, it is convenient to have a compact notation for specifying the range and interval of the argument. The notation $x = x_1(\delta x)x_2$ has come to be adopted as the standard abbreviated form for 'for values of x from x_1 to x_2 inclusive at intervals δx '. The notation mD or nS is often used for a table to m decimals, or n significant figures.

2.31. Critical tables

Most tables give the values of the function $f(x)$, rounded off to a certain number of decimals, for a sequence of equally spaced exact values of the argument x . Occasionally another type of table is more convenient, namely, one giving the *range* of x for which the function $f(x)$, rounded off to a certain number of decimals, has a specified value. Such a table is called a *critical table*, and is convenient for slowly varying functions, and also for functions which have a limited range and for which accuracy in

† By A. Fletcher, J. C. P. Miller, and L. Rosenhead (Scientific Computing Service, Ltd., London 1946). A second edition is in preparation.

the last figure is important; in using a critical table no interpolation is required, and the possibility of an error of a unit in the last figure, which occurs in interpolation in an ordinary table, is avoided.

As an example, consider a table of $\frac{1}{8}x(x-\frac{1}{2})(x-1)$, which appears in a formula for non-linear interpolation, as a function of x . A portion of a critical table of this function to four decimals is as follows:

| x | $f(x)$ |
|--------|---------|
| 0.1621 | |
| | +0.0077 |
| 0.1691 | |
| | 0.0078 |
| 0.1777 | |
| | 0.0079 |
| 0.1897 | |
| | 0.0080 |
| 0.2334 | |
| | 0.0079 |
| 0.2462 | |

The values of $f(x)$ are on lines intermediate between those on which the values of x stand, and the values of x between which a value of $f(x)$ stands mark the limits of the range of x for which that is the rounded value of $f(x)$. These values of x are rounded values of the inverse function $f^{-1}(y)$ for values of $y = f(x)$ halfway between the tabular values. For example, the above table indicates that for values of x between 0.1777 and 0.1897 the function $f(x)$ has the value +0.0079 to four decimals; and the value of x for $f(x) = 0.00785$ is 0.1777 to four decimals.

It is a convention in critical tables that if x has exactly the tabular value, the value of $f(x)$ to be taken is that standing *above* the line on which x stands; a reminder of this convention is often given in such tables by the words 'in critical cases ascend'.

2.32. Auxiliary variables in tables

An important aspect of mathematical tables is the use of auxiliary variables to simplify interpolation. This is especially important (i) in the neighbourhood of a singularity, where the ordinary interpolation formulae, applied directly to the function values, cease to be valid, (ii) for large values of the argument, and (iii) for oscillating functions when the table has to cover a large number of periods of the oscillation. Use of auxiliary variables may both simplify interpolation and lessen the amount of material which has to be calculated and printed to provide a useful table.

The most usual step is to tabulate an auxiliary function, but in some cases an auxiliary independent variable may be used instead or in

addition. The following are examples of the tabulation of auxiliary functions:

- (i) $\log\{(\sin x)/x\}$ and $\log\{(\tan x)/x\}$ for small x , in place of $\log \sin x$ and $\log \tan x$ which are infinite at $x = 0$ and cannot be interpolated by standard formulae near $x = 0$.
- (ii) If $f(x)$ is oscillatory, it may be possible to determine an 'amplitude function' $A(x)$ and a 'phase function' $\phi(x)$ such that

$$f(x) = A(x)\cos\{\phi(x)\}$$

and that $A(x)$ and $\phi'(x)$ vary much more slowly than $f(x)$. Then $A(x)$ and $\phi(x)$ can be tabulated at wider intervals than $f(x)$, and interpolation of $A(x)$ and $\phi(x)$ is easier than that of $f(x)$ itself. This is particularly convenient when two functions can be expressed as the real and imaginary parts of $A(x)\exp\{i\phi(x)\}$. An important example is provided by the Bessel functions, for which

$$A(x) = x^{\frac{1}{2}}[\{J_n(x)\}^2 + \{Y_n(x)\}^2]^{\frac{1}{2}} = x^{\frac{1}{2}}|H_n^{(1)}(x)|$$

$$\text{and} \quad \phi(x) = \tan^{-1}[Y_n(x)/J_n(x)] = \arg[H_n^{(1)}(x)]$$

form a convenient pair of auxiliary functions except for small values of x .

An example of the joint use of auxiliary functions and an auxiliary independent variable is provided by the elliptic integral

$$K(k) = \int_0^{\pi} (1 - k^2 \sin^2 \theta)^{-\frac{1}{2}} d\theta$$

near $k = 1$. If $k' = (1 - k^2)^{\frac{1}{2}}$ and

$$K(k) = K_1 \log(4/k') + K_2,$$

K_1 and K_2 are regular functions of k' near $k = 1$, and a convenient tabulation is K_1 and K_2 against k' as argument.

2.4. Slide rule

A slide rule is an instrument of limited accuracy, and of limited scope since it cannot easily be used for addition and subtraction; but within its limitations it is a valuable tool of numerical work. Two contexts in which it is particularly useful are the following:

- (i) When a function is tabulated at intervals too large for linear interpolation between tabular values, more elaborate interpolation formulae have to be used; these will be considered in Chapter V. In many of these, the interpolated value is expressed as the sum of the value which would be obtained by linear interpolation, and

some other terms which can be regarded as contribution to a 'correction' to this value. For some or all of these contributions, the accuracy attainable with a slide-rule may be adequate, and then it is a useful tool.

- (ii) When in the solution of a linear differential equation a particular integral P and a complementary function C have been evaluated, and a small constant multiple of C , say γC , has to be added to P to give a solution satisfying specified conditions, the calculation of γC may often be carried out to adequate accuracy on a slide-rule. This is a quick calculation, because after a single setting of γ , all the values of γC can be read off without resetting.

As well as the usual straight slide-rule with a 10-inch scale, there is another form, with two cursors and a single scale in the form of a helix on a movable cylinder. In the Fuller slide-rule this scale is 50 feet long, and this enables an accuracy of 1 in 10,000 to be obtained without difficulty, and 1 in 20,000 with care. Such an instrument is cheap compared with a desk machine and may be found very useful in work for which its accuracy is adequate and in circumstances in which the cost of a desk machine is prohibitive. With one of these slide-rules and an adding machine much useful numerical work can be done, especially in contexts involving empirical or experimentally determined functions not specified to more than four- or five-figure accuracy.

2.5. Graph paper

Graph paper is more generally useful as a means of presenting results than as a tool for obtaining them. But there are occasions when it is useful as a means of doing calculations, e.g. for obtaining approximate results which can later be refined by more accurate methods.

Before being used for anything more than qualitative or the roughest of quantitative work, graph paper should be examined for uniformity of ruling. Paper ruled in two colours (e.g. blue for the main ruling, with red for every tenth line) should be examined for the registration of the two colours. Paper which is ruled with every fifth or tenth line thick should be examined to see that the intervals between the *centres* of the lines are uniform, and not the intervals between the *edges* of the lines, a remarkable fault in some papers.†

2.6. Other machines

There are other aids to numerical work of various kinds, but mostly large or special pieces of equipment which are unlikely to be available

† See Jeffreys and Jeffreys, *Methods of Mathematical Physics*, chap. 9.

to most of those for whom this book is primarily intended. The more important may, however, be mentioned here.

First, there is the 'National' machine,† developed from an accounting machine. This is an adding machine with keyboard setting mechanism and six registers, with facilities for adding or subtracting the number set on the keyboard, or the number standing in any register, into any combination of registers. The mechanical arrangement for controlling these transfers is such that it can only be used effectively in calculations in which the same set of operations has to be repeated successively on different sets of numbers; but many calculations have just this character, and for such calculations this machine can be very valuable.

Secondly, there are two groups of machines for carrying out arithmetical operations on numbers represented by punchings on cards, the 'Hollerith' and 'Powers-Samas' machines. The main machines of each group are a 'tabulator' which is a multi-register adding machine with printing mechanism, a sorter, and a multiplying punch which can take a card with two numbers punched on it, and calculate and punch their product. The use of these machines, and the organization of calculations for them, is a special technique of its own,‡ and hardly appropriate for an introductory book like the present.

Thirdly, there are various high-speed automatic calculating machines which can carry out, automatically, long sequences of operations once they have been supplied with operating instructions in a suitably coded form. A short account of the principles of these machines and of the process of organizing calculations for them is given in Chapter XII.

† See, for example, L. J. Comrie, *Journ. Roy. Stat. Soc., Supplement*, **3** (1936), 87.

‡ See W. J. Eckert, *Punched Card Methods in Scientific Computation* (Columbia University, 1940).

III

EVALUATION OF FORMULAE

3.1. The significance of formulae in numerical work

THE evaluation of a given formula is the simplest kind of problem in numerical analysis. In a sense most problems reduce to this, as the numerical work itself almost always consists in substituting particular numerical values into a process or sequence of operations which could be expressed in the form of a sequence of formulae to be evaluated, even if they are not explicitly so expressed. In most cases the real question of numerical analysis is, What is the best formula or set of formulae to evaluate in order to obtain the required result?, and it is with this question that we shall primarily be concerned in later chapters. But equally important questions for practical work are how to evaluate the formulae and how to check the results.

A formula for a calculation to be carried out numerically has a significance rather different from that of a formula in formal algebra or analysis. For example, the formula

$$y = (x^2 + 1)/2x \quad (3.1)$$

regarded as an algebraical formula states a *relation* between the quantities on the two sides of the sign of equality, and is completely equivalent to

$$x^2 - 2yx + 1 = 0$$

or

$$x = y \pm (y^2 - 1)^{\frac{1}{2}}, \quad (3.2)$$

which are different ways of expressing the same relation. But formula (3.1) regarded as a formula for a numerical calculation specifies a *process* to be carried out for determining the value of y given the value of x , whereas formula (3.2) specifies a *process* to be carried out for determining the value of x given the value of y . These processes are different from one another, the data used in them are different, and the results required are different. This aspect of a formula, as representing a *process* consisting of a set of operations to be carried out in a definite sequence, plays little part in formal analysis, but is fundamental in numerical work. Even the formulae

$$x = y - z \quad \text{and} \quad z = y - x$$

mean quite different things when regarded as specifications of numerical

calculations to be carried out; and the process specified by the formula

$$y = \frac{1}{2}[x + (1/x)] \quad (3.3)$$

is different from that specified by formula (3.1).

A striking example is discussed in § 11.3, where it is shown that of two ways of writing the recurrence relation for the Bessel functions, namely,

$$J_{n+1}(x) = (2n/x)J_n(x) - J_{n-1}(x)$$

and

$$J_n(x) = (x/2n)[J_{n+1}(x) + J_{n-1}(x)],$$

which are formally completely equivalent (for $n > 0$, $x > 0$), the first specifies a numerical process which is quite impracticable as a general method for calculating $J_n(x)$ for $n > x > 0$, whereas the second gives a quite practicable iterative process.

There may be various ways of evaluating even simple formulae, and the best way may depend on the equipment available for carrying out the numerical work. For example, in the evaluation of $(abc...)/(uvw...)$ by means of a slide rule it is best to take multiplications and divisions alternately, expressed by writing this fraction in the form

$$[(a/u) \times b] / v \times c \dots$$

But with a desk machine it is best first to evaluate the denominator $D = uvw...$ and record this, then form the continued product $abc...$, and finally divide the result by D . In forming these continued products, no intermediate results need be written down; the only numbers to be recorded are D and the final result.

In using a machine it is worth while planning the calculation in such a way that as much as possible of the work is done on the machine without recording intermediate results, so as to reduce the amount of writing, with the possibilities of mistakes in recording and reading the written results, to a minimum demanded by the need for clarity in presentation of the calculation and for checking. Transfers from the counting register to the setting levers or keys should also be avoided if possible.

For example, if e^x were given, $2(\cosh x - 1)$ could be calculated from

$$2(\cosh x - 1) = e^x + (1/e^x) - 2;$$

this would require a reciprocal to be calculated, recorded, and reset on the machine (or at least transferred from the counting register to the setting levers or keys). But if it is calculated from

$$2(\cosh x - 1) = (e^x - 1)^2 / e^x,$$

this can all be done by a sequence of operations on the machine alone;

it also has the advantage that for small x it does not calculate the result as the small difference of two relatively large quantities.

3.2. Evaluation of polynomials

Expressions consisting of a number of additions and multiplications can usually be evaluated in various ways, of which the best to use in any case may depend on particular features of that case. For example, a polynomial

$$y = a_0 x^n + a_1 x^{n-1} + \dots + a_{n-1} x + a_n \quad (3.4)$$

may be evaluated by calculating the separate terms and adding. When x has a simple numerical value ($x = 1, 2$, or 10 for example) this may be the best method, especially if the coefficients are small integers. If this method is used for evaluating a polynomial both for positive and for negative values of x , a convenient procedure is first to sum separately all the terms involving odd powers of x and all those involving even powers of x , for positive values of x only, then for each value of x to add and subtract these two sums.

If x has not a simple numerical value it may be better to write

$$y = [(a_0 x + a_1)x + a_2]x + a_3]x + \dots \quad (3.5)$$

and carry out an addition and a multiplication alternately as indicated by this expression. That is, construct the sequence y_j defined by

$$y_0 = a_0, \quad y_j = y_{j-1}x + a_j \quad (j > 0); \quad (3.6)$$

the result required is y_n . This process requires n multiplications and n additions, and *no* recording of intermediate results. Care is necessary with the decimal point; use of the decimal point markers is a great help here.

The process for checking the results will depend on the calculation of which the evaluation of the polynomial forms part. It is unlikely that just a single value of a polynomial will be wanted; the evaluation of the polynomial is much more likely to form part of a larger calculation, which may well include means of checking the value obtained for the polynomial.

If a set of values y of a polynomial (3.4) for a set of values of x is calculated, then

$$\sum y = a_0 \sum x^n + a_1 \sum x^{n-1} + \dots \quad (3.7)$$

where the sum is over all values of x for which the polynomial has been calculated. One way of checking such a set of values of y is to evaluate the right-hand side of (3.7) and compare the result with $\sum y$; the results should not differ by more than the tolerance for rounding errors.

We shall see later (§ 4.42) that if a polynomial has simple coefficients and is of not too high order, its values for a set of equally-spaced values of x can be obtained simply and conveniently by a sequence of additions, without any multiplication at all.

3.3. Evaluation of power series

To evaluate the sum of a power series

$$y = a_0 + a_1 x + a_2 x^2 + \dots \quad (3.8)$$

it is often most convenient to write each term as a multiple of the preceding one, thus:

$$y = a_0 + \left(\frac{a_1}{a_0}x\right)a_0 + \left(\frac{a_2}{a_1}x\right)a_1 x + \left(\frac{a_3}{a_2}x\right)a_2 x^2 + \dots$$

and to evaluate each term from the previous one by the appropriate multiplications. Series containing odd powers only or even powers only can be treated similarly. If several values of y , at equal intervals of x , are calculated, evaluation of the finite differences (§ 4.2) of the values of y probably provides the best check.

Example: To evaluate

$$y = \frac{1}{2}x^2 + \frac{1}{2 \cdot 4 \cdot 5}x^5 + \frac{1}{2 \cdot 4 \cdot 5 \cdot 7 \cdot 8}x^8 + \frac{1}{2 \cdot 4 \cdot 5 \cdot 7 \cdot 8 \cdot 10 \cdot 11}x^{11} + \dots$$

to six decimals for $x = 1.0(0.1)1.4$.

It is convenient to write this

$$y = \frac{1}{2}x^2 \left[1 + \frac{1}{4 \cdot 5}x^3 + \frac{1}{4 \cdot 5 \cdot 7 \cdot 8}x^6 + \frac{1}{4 \cdot 5 \cdot 7 \cdot 8 \cdot 10 \cdot 11}x^9 + \dots \right] \quad (3.9)$$

and first to sum the series in the square bracket and then multiply the sum by $\frac{1}{2}x^2$.

If the ratios of successive coefficients in the series are written b_1, b_2, \dots then

$$(nth \text{ term}) = b_n x^3 [(n-1)th \text{ term}]; \quad (3.10)$$

the values of the first few b 's are

$$b_1 = \frac{1}{4 \cdot 5} = \frac{1}{20}, \quad b_2 = \frac{1}{7 \cdot 8} = \frac{1}{56}, \quad b_3 = \frac{1}{10 \cdot 11} = \frac{1}{110}, \dots,$$

and in general

$$b_n = \frac{1}{(3n+1)(3n+2)}.$$

The denominator in this fraction is a quadratic function of n , hence the second differences (see § 4.2) of its values are constant, and this can be used to check these values

| | | | | | |
|----|----|-----|-----|-----|-----|
| 20 | 56 | 110 | 182 | 272 | 380 |
| 36 | 54 | 72 | 90 | 108 | |
| 18 | 18 | 18 | 18 | 18 | |

A similar check can usually be applied if the ratios of successive coefficients can be

expressed as the ratio of two polynomials of low degree in n . The work can conveniently be arranged as follows:

| x | . | . | . | 1.0 | 1.1 | 1.2 | 1.3 | 1.4 |
|--|---|---|---|-------------|---------------------|---------------------|---------------------|-------------|
| x^3 | . | . | . | 1.000 | 1.331 | 1.728 | 2.197 | 2.744 |
| b_n | | | | 1.000000,00 | 1.000000,00 | 1.000000,00 | 1.000000,00 | 1.000000,00 |
| $1/20 = .05$ | | | | 0.050000,00 | 0.066550,00 | 0.086400,00 | 0.109850,00 | 0.137200,00 |
| $1/56 = .01785714$ | | | | 892,86 | 1581,75 | 2666,06 | 4309,65 | 6722,80 |
| $1/110 = .00909091$ | | | | 8,12 | 19,14 | 41,88 | 86,08 | 167,70 |
| $1/182 = .005495$ | | | | 0,04 | 0,14 | 0,40 | 1,04 | 2,53 |
| $1/272 = .003676$ | | | | 0,00 | 0,00 | 0,00 | 0,01 | 0,03 |
| sum | | | | 1.050901,02 | 1.068151,03 | 1.089108,34 | 1.114246,78 | 1.144093,06 |
| $\frac{1}{2}x^2$ | | | | 0.5 | 0.605 | 0.72 | 0.845 | 0.98 |
| y (to six decimals) | | | | 0.525451 | 0.646231 | 0.784158 | 0.941539 | 1.121211 |
| $\delta^2 y$ | | | | | 17147 | 19454 | 22291 | |
| $y'' = 1 + xy$ | | | | 1.52545 | 1.71085 | 1.94099 | 2.22400 | 2.56970 |
| $\delta^2 y''$ | | | | | 4474 | 5287 | 6269 | |
| $\delta^4 y''$ | | | | | | 169 | | |
| $y'' + \frac{1}{12}\delta^2 y'' - \frac{1}{240}\delta^4 y''$ | | | | | 1.7145 ₈ | 1.9453 ₉ | 2.2292 ₂ | |

Notes: (i) The entries in the third to eighth lines are the values of the terms in the square bracket in formula (3.9). Each is calculated from the preceding one by formula (3.10); if the decimal values of b_n given on the left are used, these terms can be calculated entirely by multiplication and transfer.

(ii) To obtain six decimals in the final result, it is advisable to keep eight decimals in the individual terms, that is, to retain two guarding figures.

(iii) The function y defined by the series (3.9) satisfies the equation $y'' = 1 + xy$. The second differences (see § 4.2) of y can be calculated from the values of y (see § 4.45) and compared with the values calculated from y'' by formula (4.19); this provides a close check on the results.

3.4. Kinds of formulae to avoid

There are two kinds of formulae to be avoided if possible, namely those that express the result required as

- (i) the ratio of two small numbers,
- (ii) the difference of two large, nearly equal, numbers.

When one or other of these situations occurs, it often, though not always, means that the method adopted for calculating the result is not the most suitable, and it is usually worth examining whether there is a more suitable alternative.

The following are some examples:

(a) Exponential extrapolation

Three numbers y_0 , y_1 , and y_2 are known to differ from the required result Y by amounts which are in geometrical progression; to find Y (see Fig. 5). This process is called 'exponential extrapolation'; it is useful in some methods of successive approximation (see § 9.32).

Since $y_0 - Y$, $y_1 - Y$, and $y_2 - Y$ are in geometrical progression,

$$(y_2 - Y)/(y_1 - Y) = (y_1 - Y)/(y_0 - Y)$$

and solution for Y gives

$$Y = (y_0 y_2 - y_1^2)/(y_2 - 2y_1 + y_0). \quad (3.11)$$

But if $y_0 = y_1 = y_2$ this gives $Y = 0/0$ which is useless for numerical work; and if y_0 , y_1 , and y_2 are only slightly different from Y , it gives Y as the ratio of two small numbers, the numerator and denominator being both of order $(y_0 - Y)$.

But if Y is written as the best approximation y_2 plus a correction, thus:

$$Y = y_2 - (y_2 - y_1)^2/(y_2 - 2y_1 + y_0), \quad (3.12)$$

the numerator of the 'correction' is of order $(y_0 - Y)^2$ whereas the denominator is of order $(y_0 - Y)$; the correction is therefore of order $(y_0 - Y)$ and is zero in the case $y_0 = y_1 = y_2$, and is small if y_0 is nearly equal to Y . Its evaluation gives no trouble.

This illustrates the way in which two expressions, formally equivalent, may be very different when assessed from the point of view of the ease of practical numerical evaluation.

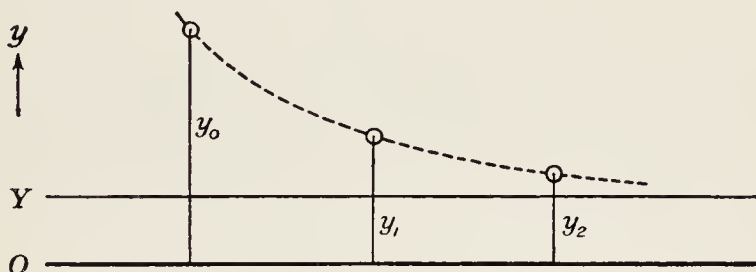


FIG. 5

(b) *Solution of a quadratic equation when the ratio of the roots is large*

Let x_1 be the larger and x_2 the smaller of the roots of the equation

$$x^2 - 18x + 1 = 0.$$

Use of the standard general formula for the root gives

$$x_1, x_2 = 9 \pm \sqrt{80}. \quad (3.13)$$

If $\sqrt{80}$ is taken to four decimals (five figures) this gives

$$x_1, x_2 = 17.9443, 0.0557.$$

Here x_2 is obtained as the small difference of two relatively large numbers 9 and $\sqrt{80} = 8.9443$; the first two significant figures in the value of $\sqrt{80}$ are lost, and from a five-figure value only a three-figure result is obtained.

On the other hand if x_2 is obtained not from (3.13) but from the relation (for this equation) $x_1 x_2 = 1$, the value of x_2 is obtained to full five-figure accuracy without requiring that $\sqrt{80}$ should be obtained to any greater accuracy than for x_1 . Here again we see a marked difference, from the point of view of numerical evaluation, between two formally equivalent formulae.

3.5. Evaluation of a function in the neighbourhood of a value of the argument at which it becomes indeterminate

In the neighbourhood of a value of the argument at which a function becomes indeterminate, some form of series expansion will usually be available.

Consider, for example, the function y defined by

$$\begin{cases} y = (1/\sin x) - (1/x) & (0 < |x| < \pi) \\ y(0) = 0. \end{cases}$$

To evaluate this for small values of x , it is convenient to write it

$$\begin{aligned} y &= x \frac{x - \sin x}{x^3} \bigg/ \frac{\sin x}{x} \\ &= \frac{x}{6} \left[1 - \frac{x^2}{4 \cdot 5} + \frac{x^4}{4 \cdot 5 \cdot 6 \cdot 7} \dots \right] \bigg/ \left[1 - \frac{x^2}{2 \cdot 3} + \frac{x^4}{2 \cdot 3 \cdot 4 \cdot 5} \dots \right]. \end{aligned}$$

It would be possible to carry out the division of one series by the other algebraically, but if more than the first two or three terms have to be included, it is easier to evaluate the two series separately and carry out the division numerically.

IV

FINITE DIFFERENCES

4.1. Functions of a continuous variable in numerical analysis

IN numerical work we may be concerned with two different ways of specifying functions of a continuous variable. First, a function may be specified by a formula which can, in principle, be evaluated for any value of x as required: examples of such functions are polynomials, circular, exponential, and other functions defined or expressed in terms of convergent power series, and functions defined by definite integrals such as the gamma-function

$$\Gamma(x) = \int_0^{\infty} e^{-t} t^{x-1} dt. \quad (4.1)$$

Secondly, there are those functions which are specified only by tables of values; these may often be tables expressing some empirical physical relationship, such as the relation between grid voltage and anode current in an electronic valve, or between velocity and resistance for a projectile; or they may be results of previous calculations.

In practice, there is not much difference between functions specified in these two ways, for usually one obtains values of functions of the first kind from tables rather than by evaluating the defining formulae. In fact, mathematical tables are made precisely for the purpose of enabling function values to be determined without going back to first principles and evaluating the defining formulae each time a function value is required; if we require $\Gamma(1.27836)$ we interpolate in tables of $\Gamma(x)$, rather than evaluate the integral in formula (4.1) for $x = 1.27836$, unless it happens that no tables to the number of figures required are available. Thus in either case we are concerned in practice with functions specified by tables, and with the properties of functions so specified.

A function $f(x)$ specified only at discrete tabular values of the independent variable x is not formally defined for intermediate values. If the tabular values of x include zero and are at equal intervals δx , and $g(x)$ is any function (not necessarily even continuous) which is finite at the tabular values of x , then $f(x) + g(x)\sin(\pi x/\delta x)$ has the same values as $f(x)$ at the tabular values of x . Further, the tabular values of $f(x)$ are usually subject to rounding errors, so that the function may not be accurately defined even at the tabular values of x .

On the other hand, a table of a function of a continuous variable x

would often be of little value unless it were possible to determine values of the function for values of x between the tabular values (to an approximation depending, of course, on rounding errors). In order to do this, some understanding is necessary about the behaviour of the function between its tabular values, an understanding which may be justified formally in cases of functions of the first kind mentioned at the beginning of this section, but may have to remain an assumption in the case of empirical functions. This understanding may be expressed qualitatively by saying that the function is 'smooth' over the range concerned. 'Smoothness' of a function is a property which it is difficult to define in a quantitative way; it is discussed further in § 11.4. It implies differentiability to some high order, and smallness of high-order derivatives. An example will illustrate this.

We shall later (Chapter V) derive interpolation formulae for use when the interval of tabulation is too large for linear interpolation between tabular values of the function. It will be found that it is possible to interpolate $\sin x$, not only roughly but to any required accuracy, from its values at interval $\delta x = \frac{1}{2}\pi$:

$$\begin{array}{cccccccc} x & -\frac{3}{2}\pi & -\pi & -\frac{1}{2}\pi & 0 & +\frac{1}{2}\pi & \pi & \frac{3}{2}\pi \\ y = \sin x & 1 & 0 & -1 & 0 & 1 & 0 & -1 \end{array} \quad (4.2)$$

or even from its values at intervals $x = \frac{2}{3}\pi$.

Let us inquire what particular property the function $y = \sin x$ has which selects it from all other functions with the tabular values (4.2) as the one for which the interpolation formula is accurate. Suppose, for simplicity, that we know that y is an odd function of x , periodic with period 2π . Then it can be expanded in a sine series in the interval $-\pi \leq x \leq \pi$:

$$y = a_1 \sin x + a_3 \sin 3x + a_5 \sin 5x + \dots \quad (4.3)$$

where, to give the value of y at $x = \frac{1}{2}\pi$,

$$1 = a_1 - a_3 + a_5 - a_7 + \dots \quad (4.4)$$

We shall require a measure of the n th derivative of y ; this derivative varies with x , but a convenient single quantity giving an overall measure of its magnitude is its mean square value

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} (y^{(n)})^2 dx = \frac{1}{2}(a_1^2 + 2^n a_2^2 + 3^n a_3^2 + \dots).$$

As $n \rightarrow \infty$, the minimum value of this quantity, subject to the condition (4.4), is given by $a_1 = 1$, $a_m = 0$ ($m > 1$). Thus the relevant property of the function $y = \sin x$ is this, that of all functions which are odd and have period 2π , it is the one for which, to put it roughly, the high-order derivatives are as small as possible.

Another aspect of this property of the function $\sin x$, that it can be interpolated accurately from its values at a wide interval such as $\frac{1}{2}\pi$ or $\frac{2}{3}\pi$, is considered in § 5.91.

In the great majority of cases, functions are tabulated at equal intervals of the independent variable, which is often called, in this

context, the 'argument' of the table. For the present, we will only consider such sets of function values.

4.2. Finite differences

The most important property of a function specified by a table consists of what are called its 'finite differences'. The following example illustrates what is meant by this term:

| x | $f(x) = 1/x$ | <i>First differences</i> | <i>Second differences</i> | <i>Third differences</i> |
|-----|--------------|------------------------------|-------------------------------|------------------------------|
| 3.0 | 0.33333 | | | |
| | | — 1075 | | |
| 3.1 | .32258 | | 67 | |
| | | — 1008 | | — 6 |
| 3.2 | .31250 | | 61 | |
| | | — 947 | | — 5 |
| 3.3 | .30303 | | 56 | |
| | | — 891 | | — 6 |
| 3.4 | .29412 | | 50 | |
| | | — 841 | | — 2 |
| 3.5 | .28571 | | 48 | |
| | | — 793 | | |
| 3.6 | 0.27778 | | | |

The 'first differences' are obtained by subtracting each function value from that for the next *greater* tabular value of x ; the 'second differences' are obtained by carrying out a similar set of subtractions on the first differences, and so on.

Values of *odd* order differences should be written on levels intermediate between those of function values, and values of *even* order differences on the same lines as function values; normally these values are written in terms of the last digital position as unit, decimal points, and zeros before the first significant digit, being omitted. They can conveniently be distinguished from function values by being written or printed smaller.

The finite differences of tabular functions play a very important part both in the analytical and in the numerical manipulation of such functions. Use of them enables formulae for operations on such functions, such as interpolation and integration, to be expressed compactly and in a form convenient for practical use. When the tabular values provide all the information we have about a function, all processes involving this function have to be expressed as operations on the tabular values; one of the most important operations on a set of values at equal intervals of the independent variable is that of differencing, and we shall see, later in this Chapter and in Chapters V and VI, that most of the other operations can be expressed in terms of this one.

It will be seen that the values of the third differences in the above table are noticeably irregular; this is an effect of rounding errors in the function value, which will be considered more fully in § 4.44.

In order that a function shall be well determined by a table, the average value of the n th order differences should tend to zero, or at least become small, as n increases. We have seen examples in which this does not appear to be the case; for the function $y = \sin x$, at interval $\frac{1}{2}\pi$ in x (see (4.2)), the $2n$ th differences have extreme values $\pm 2^n$, but this function is still well defined by these values, in the sense that accurate intermediate values can be interpolated between them. But this is a peculiar property of the function $y = \sin x$ alone out of all the functions with these tabular values; given these function values *alone*, without the knowledge that they are intended to represent $\sin x$, one could not be at all confident about the results of any attempt to interpolate between them. Eight values per cycle is about the smallest number which can in practice be regarded as specifying an oscillating function adequately, and at least twelve values per cycle is preferable.

4.21. Notation for finite differences

Let x_0 be one of the tabular values of x , $x_j = (x_0 + j\delta x)$ a set of other tabular values, and $f_j = f(x_j)$ the values of $f(x)$ at the tabular values of x .

There are two kinds of notation for finite differences. In one the differences of a function f are written δf or Δf , so that the symbol δ or Δ stands for an *operation* carried out on the values of the function f . In the other the symbol δ or Δ is used for the *differences themselves*.

The former seems much the preferable, both for use in the derivation and manipulation of formulae in finite differences and in application of them. It is more nearly self-explanatory, and many formulae with which we shall be concerned express relations between differences of two different functions (for example a function of x and its derivative), and if a symbol is used to represent a difference itself rather than a difference-operator, differences of different functions cannot be distinguished except by introducing new symbols, which are unnecessary. In this notation, repetition of an operation is expressed by the use of an index (as in δ^2 , Δ^3),

The use of the symbol δ or Δ for the differences themselves is a convenient shorthand in cases in which it is unambiguous, and is sometimes preferred by those carrying out the details of the numerical work. In this notation the use of dashes (Δ'') or Roman superiors (as $\Delta^{(iv)}$) is preferable to the use of numerical indices to indicate orders of differences.

In this book, the former usage will be adopted throughout, so that δ

and Δ must be regarded as finite-difference *operators*. Consistently with this notation, δx will be used for the interval in the independent variable (other notations for this are h and w).[†]

The first difference $f_1 - f_0$ may be associated with the argument value x_0 , with the argument value x_1 , or symmetrically with these two argument values, and assigned a corresponding suffix. A different symbol for the finite-difference operator is used to distinguish these three cases:

$$\left. \begin{aligned} f_1 - f_0 &= \Delta f_0 \\ &= \nabla f_1 \\ &= \delta f_1 \end{aligned} \right\}. \quad (4.5)$$

This is generalized in the following three schemes for a difference table:

| x f | | | | f | | | | f | | | |
|-----------------|-----------------|-------------------|-------------------|----------|-----------------|-------------------|----------------|----------|-------------------|----------------------------|------------------------------|
| $x_{-2} f_{-2}$ | | $\Delta^2 f_{-2}$ | | f_{-2} | | $\nabla^2 f_{-1}$ | | f_{-2} | | $\delta^2 f_{-1}$ | |
| | Δf_{-1} | | $\Delta^3 f_{-1}$ | | ∇f_{-1} | | $\nabla^3 f_0$ | | | $\delta f_{-1\frac{1}{2}}$ | |
| $x_{-1} f_{-1}$ | | $\Delta^2 f_{-1}$ | | f_{-1} | ∇f_0 | | $\nabla^2 f_1$ | | f_{-1} | | $\delta^2 f_{-1\frac{1}{2}}$ |
| | Δf_{-1} | | $\Delta^3 f_{-1}$ | | ∇f_0 | | $\nabla^3 f_1$ | | | $\delta f_{-\frac{1}{2}}$ | |
| $x_0 f_0$ | | $\Delta^2 f_{-1}$ | | f_0 | ∇f_1 | | $\nabla^2 f_2$ | | $\underline{f_0}$ | | $\underline{\delta^2 f_0}$ |
| | Δf_0 | | $\Delta^3 f_{-1}$ | | ∇f_1 | | $\nabla^3 f_2$ | | | $\delta f_{\frac{1}{2}}$ | |
| $x_1 f_1$ | | $\Delta^2 f_0$ | | f_1 | ∇f_2 | | $\nabla^3 f_2$ | | f_1 | | $\delta^2 f_1$ |
| | Δf_1 | | $\Delta^3 f_0$ | | ∇f_2 | | $\nabla^3 f_2$ | | | $\delta f_{1\frac{1}{2}}$ | |
| $x_2 f_2$ | | $\Delta^2 f_1$ | | f_2 | | $\nabla^2 f_2$ | | f_2 | | $\delta^2 f_2$ | |

In any particular numerical case the *numbers* will be the same in each table; what is different is the general notation for these numbers, the notation which expresses the value of x with which each difference is associated.

Differences with the same suffix value in table (a) are called ‘forward differences’; they lie on a downward-slanting line on the table, such as those underlined. The forward differences from the first entry in a table are sometimes called ‘leading differences’. Those differences with the same suffix value in table (b) are called ‘backward differences’; an example is indicated similarly. Those with the same suffix in table (c) are called ‘central differences’.

Central differences are much the most useful in practice. Many formulae in central differences involve only alternate orders of differences, whereas the corresponding formulae in forward or backward differences involve all orders of differences; also the coefficients of higher terms in central-difference formulae usually decrease more rapidly with the order n of the differences than do the coefficients in formulae involving forward or backward differences. Further, this notation gives a much more natural relation between finite differences and derivatives.

† It is sometimes convenient to distinguish between the general symbol δx for the interval length and the particular value which it has in a particular calculation.

In the analytical work of deriving formulae for interpolation, integration, etc., in terms of differences, use of forward differences leads to rather simpler algebra; but in order to get from the results the central-difference formulae which are most convenient for practical use, it may be necessary to do some rather laborious algebra, which may then only give the coefficients of the central-difference formulae term by term, and be difficult to generalize to give the general term. It seems best to work throughout in terms of central differences, and so obtain directly the formulae for interpolation, integration, etc., in the forms in which they are most useful for practical work. The symbol Δ is then left free for another use, to indicate the difference between the data or between the results of two similar calculations.

It will be seen that in the central-difference scheme (c) on p. 37, only the *even*-order differences have integral suffixes. It is sometimes convenient to take the arithmetic mean of two adjacent differences and to write

$$\begin{aligned} \mu \delta f_0 &= \tfrac{1}{2}(\delta f_{\frac{1}{2}} + \delta f_{-\frac{1}{2}}), \\ \text{and in general} \quad \mu \delta^n f_j &= \tfrac{1}{2}(\delta^n f_{j+\frac{1}{2}} + \delta^n f_{j-\frac{1}{2}}) \\ &= \tfrac{1}{2}(\delta^{n-1} f_{j+1} - \delta^{n-1} f_{j-1}). \end{aligned}$$

Then the available differences are *odd*-order differences with (integer + $\frac{1}{2}$) suffixes, and *even*-order differences and *odd*-order *mean* differences with integral suffixes. A set of successive function values f_j from $j = J - k$ to $J + k$ inclusive is said to be 'centred on' the argument value x_J or on the function value f_J ; similarly for a set of differences $\delta^n f_j$.

4.3. Finite differences in terms of function values

It is sometimes convenient to have differences expressed in terms of the function values from which they are derived. We have in succession

$$\delta f_{\frac{1}{2}} = f_1 - f_0, \tag{4.6}$$

$$\begin{aligned} \delta^2 f_0 &= \delta f_{\frac{1}{2}} - \delta f_{-\frac{1}{2}} = (f_1 - f_0) - (f_0 - f_{-1}) \\ &= f_1 - 2f_0 + f_{-1}, \end{aligned} \tag{4.7}$$

$$\begin{aligned} \delta^3 f_{\frac{1}{2}} &= \delta^2 f_1 - \delta^2 f_0 = (f_2 - 2f_1 + f_0) - (f_1 - 2f_0 + f_{-1}) \\ &= f_2 - 3f_1 + 3f_0 - f_{-1}, \end{aligned}$$

$$\text{and in general} \quad \delta^n f_j = \sum_{k=0}^n (-1)^k \frac{n!}{k!(n-k)!} f_{j+\frac{1}{2}n-k} \tag{4.8}$$

as can be proved by induction; for an alternative proof see § 4.6. The coefficients of the function values in $\delta^n f_j$ are those in the binomial expansion of $(1-z)^n$.

In particular, the differences of the function

$$\begin{cases} f_m = 0 & m \neq 0 \\ f_0 = 1 \end{cases}$$

are the binomial coefficients:

| x | f | δf | $\delta^2 f$ | $\delta^3 f$ | $\delta^4 f$ |
|----------|-----|------------|--------------|--------------|--------------|
| x_{-2} | 0 | | 0 | | 1 |
| | | 0 | | 1 | |
| x_{-1} | 0 | | 1 | | -4 |
| | | 1 | | -3 | |
| x_0 | 1 | | -2 | | 6 |
| | | -1 | | 3 | |
| x_1 | 0 | | 1 | | -4 |
| | | 0 | | -1 | |
| x_2 | 0 | | 0 | | 1 |

The effect of an error ϵ in a function value on the difference table builds up in the same way:

| x | F | | | | |
|----------|------------------|--------------------------------------|------------------------------|---|-------------------------------|
| x_{-2} | f_{-2} | | $\delta^2 f_{-2}$ | | $\delta^4 f_{-2} + \epsilon$ |
| | | $\delta f_{-1\frac{1}{2}}$ | | $\delta^3 f_{-1\frac{1}{2}} + \epsilon$ | |
| x_{-1} | f_{-1} | | $\delta^2 f_{-1} + \epsilon$ | | $\delta^4 f_{-1} - 4\epsilon$ |
| | | $\delta f_{-\frac{1}{2}} + \epsilon$ | | $\delta^3 f_{-\frac{1}{2}} - 3\epsilon$ | |
| x_0 | $f_0 + \epsilon$ | | $\delta^2 f_0 - 2\epsilon$ | | $\delta^4 f_0 + 6\epsilon$ |
| | | $\delta f_{\frac{1}{2}} - \epsilon$ | | $\delta^3 f_{\frac{1}{2}} + 3\epsilon$ | |
| x_1 | f_1 | | $\delta^2 f_1 + \epsilon$ | | $\delta^4 f_1 - 4\epsilon$ |
| | | $\delta f_{1\frac{1}{2}}$ | | $\delta^3 f_{1\frac{1}{2}} - \epsilon$ | |
| x_2 | f_2 | | $\delta^2 f_2$ | | $\delta^4 f_2 + \epsilon$ |

This is the basis of an important application of differences to *checking* tables, and sometimes for correcting isolated errors, which will be considered shortly (§ 4.43).

4.4. Simple applications of differences

The simplest applications of differences are:

- (a) Building up polynomials;
- (b) Checking tables;
- (c) Smoothing.

Of these, (a) and (b) will be considered here and (c) in § 11.4.

4.41. Differences of a polynomial

An important property of finite differences is that for a polynomial of degree n , the n th order differences are constant. One proof of this is by induction.

Suppose that
$$\delta^m(x^n) = m! (\delta x)^m \quad (4.9)$$

for all integral values of m up to $m = n$, say; then it will be proved that (4.9) holds for $m = n+2$. Since for $p < m$

$$\delta^m(x^p) = \delta^{m-p}(\delta^p x^p),$$

$$(4.9) \text{ implies that } \delta^m(x^p) = 0 \quad \text{for } p < m. \quad (4.10)$$

Now from (4.7)

$$\begin{aligned} \delta^2(x^{n+2}) &= (x+\delta x)^{n+2} - 2x^{n+2} + (x-\delta x)^{n+2} \\ &= (n+2)(n+1)(\delta x)^2 x^n + \text{terms of lower degree,} \end{aligned}$$

so

$$\begin{aligned} \delta^{n+2}x^{n+2} &= \delta^n(\delta^2x^{n+2}) \\ &= (n+2)(n+1)(\delta x)^2 \delta^n[x^n + \text{terms of lower degree}] \\ &= (n+2)(n+1)(\delta x)^2 \delta^n(x^n) \quad [\text{by (4.10)}] \\ &= (n+2)(n+1)(\delta x)^2 n! (\delta x)^n \quad [\text{since (4.9) holds for } m = n] \\ &= (n+2)! (\delta x)^{n+2}. \end{aligned}$$

Now (4.9) holds for $m = 1$ and $m = 2$, hence the induction succeeds, and (4.9) holds for all integral m .

It follows that for a polynomial of degree m , say

$$p_m(x) = a_0 x^m + a_1 x^{m-1} + \dots,$$

the m th differences are constant and equal to $a_0 m! (\delta x)^m$.

Another derivation, which depends on some properties of a set of functions to which reference will be made later, is as follows. Consider the polynomials:†

$$\begin{aligned} \psi_0(\xi) &= 1, \\ \psi_m(\xi) &= \prod_{k=0}^{m-1} [\xi + \{\tfrac{1}{2}(m-1) - k\}]; \end{aligned} \quad (4.11)$$

$\psi_m(\xi)$ is a polynomial of degree m in ξ , and its argument value is the mean of the extreme factors. The first few such functions are

$$\begin{aligned} \psi_0(\xi) &= 1, \\ \psi_1(\xi) &= \xi, \\ \psi_2(\xi) &= (\xi + \tfrac{1}{2})(\xi - \tfrac{1}{2}) = (\xi^2 - \tfrac{1}{4}), \\ \psi_3(\xi) &= (\xi + 1)\xi(\xi - 1) = \xi(\xi^2 - 1), \\ \psi_4(\xi) &= (\xi + \tfrac{3}{2})(\xi + \tfrac{1}{2})(\xi - \tfrac{1}{2})(\xi - \tfrac{3}{2}) \\ &= (\xi^2 - \tfrac{1}{4})(\xi^2 - \tfrac{9}{4}). \end{aligned}$$

For intervals $\delta\xi = 1$, the first differences of ψ_m are

$$\begin{aligned} \delta\psi_m(\xi + \tfrac{1}{2}) &= \psi_m(\xi + 1) - \psi_m(\xi) \\ &= [\xi + \tfrac{1}{2}(m+1)][\xi + \tfrac{1}{2}(m-1)] \dots [\xi - \tfrac{1}{2}(m-3)] - \\ &\quad - \underbrace{[\xi + \tfrac{1}{2}(m-1)] \dots [\xi - \tfrac{1}{2}(m-3)]}_{\text{common factors}} [\xi - \tfrac{1}{2}(m-1)]. \end{aligned}$$

† Sometimes called 'factorial polynomials'.

The common factor of the two terms, indicated by a bracket, is a polynomial of the set (4.11); it has $(m-1)$ factors and the mean of its extreme factors is $(\xi + \frac{1}{2})$, so it is $\psi_{m-1}(\xi + \frac{1}{2})$. Hence, for $m > 0$,

$$\begin{aligned}\delta\psi_m(\xi + \tfrac{1}{2}) &= [\{\xi + \tfrac{1}{2}(m+1)\} - \{\xi - \tfrac{1}{2}(m-1)\}]\psi_{m-1}(\xi + \tfrac{1}{2}) \\ &= m\psi_{m-1}(\xi + \tfrac{1}{2}),\end{aligned}\tag{4.12}$$

and $\delta\psi_0 \equiv 0$.

Repeating the operation we have

$$\begin{aligned}\delta^2\psi_m(\xi) &= \delta\psi_m(\xi + \tfrac{1}{2}) - \delta\psi_m(\xi - \tfrac{1}{2}) \\ &= m[\psi_{m-1}(\xi + \tfrac{1}{2}) - \psi_{m-1}(\xi - \tfrac{1}{2})] \\ &= m\delta\psi_{m-1}(\xi) \\ &= m(m-1)\psi_{m-2}(\xi),\end{aligned}$$

and ultimately

$$\begin{aligned}\delta^m\psi_m(\xi) &= m! \\ \delta^{m+1}\psi_m(\xi) &= 0.\end{aligned}$$

Any polynomial $p_m(x) = a_0x^m + a_1x^{m-1} + \dots$, tabulated at intervals (δx) in x , can be written

$$p_m(x) = a_0(\delta x)^m[\psi_m(\xi) + b_1\psi_{m-1}(\xi) + b_2\psi_{m-2}(\xi) + \dots]$$

where $\xi = x/\delta x$, so

$$\delta^m p_m(x) = a_0 m! (\delta x)^m$$

as already shown.

This result, that the m th differences of any polynomial of degree m are constant, and its $(m+1)$ th differences are zero, corresponds, in finite differences, to the result in differential calculus that the m th derivative of such a polynomial is constant and its $(m+1)$ th derivative zero. The functions $\psi_m(x)$ take the place, in finite differences, of the functions x^n in differential calculus, as the polynomials whose form remains unchanged on differencing.

These functions will appear later in another context, for which some further properties of them will be required. From the definition (4.11) it follows that

$$\psi_m(-\xi) = (-)^m \psi_m(\xi).$$

Hence for odd values of m

$$\begin{aligned}\psi_{2n+1}(\xi) + \psi_{2n+1}(1-\xi) &= \psi_{2n+1}(\xi) - \psi_{2n+1}(\xi - 1) \\ &= (2n+1)\psi_{2n}(\xi - \tfrac{1}{2})\end{aligned}\tag{4.13}$$

by (4.12), whereas

$$\begin{aligned}\psi_{2n+1}(\xi) - \psi_{2n+1}(1-\xi) &= \psi_{2n+1}(\xi) + \psi_{2n+1}(\xi - 1) \\ &= [(\xi + n) + (\xi - n - 1)]\psi_{2n}(\xi - \tfrac{1}{2}) \\ &= (2\xi - 1)\psi_{2n}(\xi - \tfrac{1}{2}).\end{aligned}\tag{4.14}$$

4.42. Building up polynomials

The constancy of the m th differences of an m th order polynomial can be used to construct a table of values of the polynomial by building up successively the lower orders of differences from the higher by repeated addition. It is necessary to calculate at least m function values to give a set of leading differences from which to start the construction of the difference table, and it is advisable to take one or two more to provide a check.

Example: To evaluate the polynomial $y = x^3 - 5x^2 + 6x + 1$ for $x = 0(1)10$:

| x | x^3 | $-5x^2$ | $+6x$ | $+1 =$ | y | δy | $\delta^2 y$ | $\delta^3 y$ |
|-----|-------|---------|-------|--------|-----|------------|--------------|--------------|
| -2 | -8 | -20 | -12 | $+1 =$ | -39 | | | |
| -1 | -1 | -5 | -6 | $+1 =$ | -11 | +28 | -16 | |
| 0 | 0 | 0 | 0 | $+1 =$ | 1 | 12 | -10 | 6 |
| 1 | 1 | -5 | 6 | $+1 =$ | 3 | 2 | -4 | 6 |
| 2 | 8 | -20 | +12 | $+1 =$ | 1 | -2 | | 6 |
| 3 | | | | | 1 | 0 | 2 | 6 |
| 4 | | | | | 9 | 8 | 8 | 6 |
| 5 | | | | | 31 | 22 | 14 | 6 |
| 6 | | | | | 73 | 42 | 20 | 6 |
| 7 | | | | | 141 | 68 | 26 | 6 |
| 8 | | | | | 241 | 100 | 32 | 6 |
| 9 | | | | | 379 | 138 | 38 | 6 |
| 10 | 1000 | -500 | +60 | $+1 =$ | 561 | 182 | 44 | 6 |
| | | | | | | | | check |

Here five function values, from $x = -2$ to 2 (the simplest ones to evaluate) have been calculated to provide a start for building up the differences. We know from (4.9) that the third differences must have the constant value 6, and this provides a check on the starting values. From the constant third differences of 6 the second differences are built up, then the first differences, and finally the function values. The function value at $x = 10$ is easy to calculate directly, and is so calculated to provide a check on the successive additions.

It will be noted that intermediate values of y are calculated by *addition only*: this process can be carried out very effectively on an adding machine fitted with a printing mechanism (§ 2.26). For example, in summing the second differences to give the first differences, after adding each second difference the resulting value of the first difference, which is the current total, is printed without clearing by taking a ‘sub-total’. The results appear in the form of alternate values of second differences and first differences; the former can be checked against the table of values and the latter then summed similarly to give the function values.

It is necessary in using this process to keep all figures without rounding off, although final results may not be wanted to this accuracy.

Example: To evaluate the polynomial $y = x^3 - 5x^2 + 6x + 1$ for $x = 0(0.01)0.1$; four decimals required.

| x | x^3 | $-5x^2$ | $+6x$ | $+1$ | $= y$ | $\delta^3 y$ | y rounded off to 'four decimals | | |
|-------|-----------|---------|-------|------|-------------|--------------|--------------------------------------|--------|-------|
| -0.02 | -0.000,08 | -0.0020 | -0.12 | +1 | = 0.8779,92 | | 0.8780 | | |
| 0.01 | -0.000,01 | -0.0005 | -0.06 | +1 | = 0.9394,99 | 61507 | | 615 | |
| 0 | 0 | 0 | 0 | +1 | = 1.0000,00 | 60501 | -1006 | 0.9395 | - 10 |
| 0.01 | +0.000,01 | -0.0005 | +0.06 | +1 | = 1.0595,01 | 59501 | -1000 | 1.0000 | - 10 |
| 0.02 | +0.000,08 | -0.0020 | +0.12 | +1 | = 1.1180,08 | 58507 | - 994 | 1.0595 | - 10 |
| 0.03 | | | | | 1.1755,27 | 57519 | - 988 | 1.1180 | - 10 |
| 0.04 | | | | | 1.2320,64 | 56537 | - 982 | 1.1755 | - 9 |
| 0.05 | | | | | 1.2876,25 | 55561 | - 976 | 1.2321 | - 101 |
| 0.06 | | | | | 1.3422,16 | 54591 | - 970 | 1.2786 | + 171 |
| 0.07 | | | | | 1.3958,43 | 53627 | - 964 | 1.3422 | - 100 |
| 0.08 | | | | | 1.4485,12 | 52669 | - 958 | 1.3958 | - 9 |
| 0.09 | | | | | 1.5002,29 | 51717 | - 952 | 1.4485 | - 10 |
| 0.10 | | | | | 1.5510,00 | 50771 | - 946 | 1.5002 | - 9 |
| | | | | | | | | 1.5510 | |

Notes: (i) Although the third difference of 6 in the sixth decimal is smaller than the rounding error in the four-decimal values finally required, it must not be neglected on that account, as this would be a *systematic* rounding error which would accumulate and ultimately affect the results wanted. Omission of it would be equivalent to omitting the x^3 term in the polynomial, and the error would already be 10 in the fourth decimal at $x = 0.1$.

(ii) Here a typical copying mistake (78 for 87) has been made in the column of rounded-off values, which are those finally required. Such a mistake is easy to make at this stage; all the calculations have been done, and all that is wanted is to copy the four decimals required with the appropriate rounding off; unconsciously one may relax some of the care with which the rest of the calculation has been carried out, and then a mistake of this kind can easily occur. Such a mistake is easily identified by *differencing the rounded-off results* and such a check should *always* be used. As will be seen in the following section, the irregular differences not only locate the erroneous value unambiguously, but strongly suggest the correction.

4.43. Checking by differences

We have seen in § 4.3 that an isolated error ϵ in a function value makes:

| | |
|-----------------|-------------------------------------|
| a maximum error | ϵ in the first differences |
| | 2ϵ second differences |
| | 3ϵ third differences |
| | 6ϵ fourth differences |
| | 10ϵ fifth differences |
| | 20ϵ sixth differences |

whereas the magnitude of the differences themselves normally decreases with the order of differences; if it does not, the function is not well defined by the table. Hence an error shows up more and more as the order of the differences is increased. Examination of the differences of a function is one of the best checks against *random* errors; it will not necessarily check against *systematic* errors.

The differences which are affected by an error spread fanwise from the incorrect function value (see § 4.3), and this can be used to locate an error.

Example:

| x | y | $\delta^2 y$ | | | $\delta^4 y$ | corrections to $\delta^4 y$ |
|-----|------|--------------|-----|-----|--------------|--------------------------------|
| 0 | 358 | | | | | |
| | | 12 | | | | |
| 1 | 370 | | 15 | | | |
| | | 27 | | 12 | | |
| 2 | 397 | | 27 | | - 1 | |
| | | 54 | | 11 | | |
| 3 | 451 | | 38 | | - 1 | |
| | | 92 | | 10 | | |
| 4 | 543 | | 48 | | - 1 | |
| | | 140 | | 9 | | |
| 5 | 683 | | 57 | | - 19 | + 18 |
| | | 197 | | -10 | | |
| 6 | 880 | | 47 | | + 71 | - 72 |
| | | 244 | | +61 | | |
| 7 | 1124 | | 108 | | -109 | +108 |
| | | 352 | | -48 | | |
| 8 | 1476 | | 60 | | + 71 | - 72 |
| | | 412 | | 23 | | |
| 9 | 1888 | | 83 | | - 19 | + 18 |
| | | 495 | | 4 | | |
| 10 | 2383 | | 87 | | - 1 | |
| | | 582 | | 3 | | |
| 11 | 2965 | | 90 | | | |
| | | 672 | | | | |
| 12 | 3637 | | | | | |

The last column is $18 \times (1, -4, 6, -4, 1)$.

Notes: (i) The existence *and location* of an error is unambiguously shown by the table.

(ii) A change Δy in a function value y makes changes $(1, -4, 6, -4, 1)$ times Δy in successive values of the fourth difference, centred on the changed value of y . A few trials show that a change $\Delta y = +18$ will make all the fourth differences -1 . The error can often be corrected in this way.

(iii) A transposition of two adjacent digits differing by m will produce an error of $9m$ in terms of the less significant of the digits as unit. It has already been mentioned that transpositions form a common type of mistake; values of Δy which are multiples of 9, or nearly, probably arise from mistakes of this kind. This can be checked from the values of the digits involved. Here $\Delta y = 18$, hence $m = 2$ in the last figure. The value $y = 1124$ at $x = 7$ should read $y = 1142$.

(iv) In this case the fourth difference of the corrected table is exact; the location and correction of the mistake is not affected by rounding errors.

Example: Here the values of y are alleged to be rounded off from a table of $x^{\frac{1}{2}}$:

| x | y | δy | $\delta^2 y$ | Correction to $\delta^2 y$ | Revised $\delta^2 y$ | Correction to $\delta^2 y$ | Corrected $\delta^2 y$ |
|-----|--------|------------|--------------|-------------------------------|-------------------------|-------------------------------|---------------------------|
| 38 | 6.1644 | 806 | | | | | |
| 39 | .2450 | 796 | -10 | | -10 | | -10 |
| 40 | .3246 | 785 | -11 | | -11 | | -11 |
| 41 | .4031 | 776 | -9 | | -9 | | -9 |
| 42 | .4807 | 740 | -36 | +27 | -9 | | -9 |
| 43 | .5547 | 785 | +45 | -54 | -9 | | -9 |
| 44 | .6332 | 750 | -35 | +27 | -8 | | -8 |
| 45 | .7082 | 750 | 0 | -9 | -9 | | -9 |
| 46 | .7832 | 745 | -5 | +18 | +13 | -20 | -7 |
| 47 | .8577 | 705 | -40 | -9 | -49 | +40 | -9 |
| 48 | 6.9282 | 718 | +13 | | +13 | -20 | -7 |
| 49 | 7.0000 | 711 | -7 | | -7 | | -7 |
| 50 | .0711 | 703 | -8 | | -8 | | -8 |
| 51 | .1414 | 697 | -6 | | -6 | | -6 |
| 52 | .2111 | | | | | | |

Notes: (i) Here a succession of seven values of $\delta^2 y$ is irregular. The first obviously wrong value, -36 at $x = 42$, indicates a mistake at $x = 43$; the value of $\delta^2 y(42)$ would be expected to be -8 , -9 , or -10 ; that is, the correction is $+28$, $+27$, or $+26$. The value $+27$ suggests a transposition of two digits differing by 3 in $y(43)$, and reference to the function values shows that the end digits *do* differ by 3.

(ii) Correction of this mistake then makes the differences smooth, apart from slight irregularities which can be ascribed to rounding errors in the function values, as far as $\delta^2 y(44)$ inclusive. The next *four* second differences are irregular, indicating mistakes in both the values $y(46)$ and $y(47)$. The value of $\delta^2 y(45)$ would be expected to be -8 or -9 ; that is the correction is -8 or -9 ; the latter suggests an interchange of two digits differing by 1, and when the corresponding correction has been made, we have the series of second differences given in the column headed 'Revised'.

(iii) The next three second differences should be about -8 , and to give them all this value we would require corrections $(-21, +41, -21)$; the corrections arising from a single change in y must be in the ratio $(1:-2:1)$, so the error in $y(47)$ is $+20$ or $+21$. The former would be produced by doubling the wrong one of two digits differing by 2, and as such digits do occur in $y(47)$ in the right place, the error can be ascribed to this cause with fair certainty.

These examples show that it is possible to use differences not only for detection and location of errors in tables, but for correcting them, when the nature of the error is clear from the behaviour of the differences, or for indicating a probable correction when it is not. In the case of the second example just given, it would of course be much better to use the differences simply to *indicate* the erroneous values, and to refer back to a table of $x^{\frac{1}{2}}$ to *correct* them.

4.44. Effect of rounding errors on differences

In most tables almost every function value will be in error to some degree, on account of rounding errors. Although the rounding error in a function value may not be more than $\frac{1}{2}$ in the least significant figure, the effect of an error is exaggerated in the higher differences, which unavoidably become somewhat irregular, and the more so the higher the order of differences. It is important to realize this, otherwise irregularities in differences which are due to rounding errors may be taken as indicating mistakes, and time may be spent trying to find mistakes and to make changes in function values which cannot be improved except by taking more significant figures.

The greatest effects of rounding errors will occur when alternate function values are rounded off by $+\frac{1}{2}$ and $-\frac{1}{2}$ alternately. Then departures of the n th differences from those for unrounded function values may be up to $\pm 2^{n-1}$ in the last place tabulated and alternate departures will be of alternate signs; though such large irregularities will be rarer the higher the order n of the differences. It is useful to have a working criterion for the magnitude of the fluctuations in the different orders of differences which can be expected as the result of rounding errors. Comrie† gives the following limits for various values of n :

| n | 1 | 2 | 3 | 4 | 5 | 6 | 8 | 10 |
|-----|---------|---------|---------|---------|----------|----------|----------|-----------|
| | ± 1 | ± 2 | ± 3 | ± 6 | ± 12 | ± 22 | ± 80 | ± 300 |

Differences having fluctuations less than these limits can be accepted; only those having greater fluctuations should be regarded as suspicious.

The example opposite illustrates the way in which irregular differences may occur in the most accurate rounded-off values of a smooth function.

From examination of the differences in the table, one would be very inclined to 'correct' the rounded values of $y(5)$ and $y(8)$ to 393 and 777 respectively, giving the third differences shown on the extreme right; but reference to the exact values of y shows that this would be incorrect.

† *Chambers's 6-Figure Mathematical Tables*, vol. 2 (1949), Introduction, p. xxxi.

| x | y | δy | $\delta^2 y$ | $\delta^3 y$ | $\delta^4 y$ | Values of y rounded off to nearest unit | | | | 'Correc- tions' | |
|-----|---------|------------|--------------|--------------|--------------|--|------------|--------------|--------------|--------------------|--------------|
| | | | | | | y | δy | $\delta^2 y$ | $\delta^3 y$ | to $\delta^3 y$ | $\delta^3 y$ |
| 0 | 61.24 | | | | | 61 | | | | | |
| | | 4581 | | | | | 46 | | | | |
| 1 | 107.05 | | 897 | | | 107 | | 9 | | | |
| | | 5478 | | 109 | | | 55 | | 1 | | 1 |
| 2 | 161.83 | | 1006 | | 53 | 162 | | 10 | | | |
| | | 6484 | | 162 | | | 65 | | 1 | | 1 |
| 3 | 226.67 | | 1168 | | 52 | 227 | | 11 | | | |
| | | 7652 | | 214 | | | 76 | | 4 | -1 | 3 |
| 4 | 303.19 | | 1382 | | 51 | 303 | | 15 | | | |
| | | 9034 | | 265 | | | 91 | | 0 | +3 | 3 |
| 5 | 393.53 | | 1647 | | 50 | 394 | | 15 | | | |
| | | 10681 | | 315 | | | 106 | | 6 | -3 | 3 |
| 6 | 500.34 | | 1962 | | 49 | 500 | | 21 | | | |
| | | 12643 | | 364 | | | 127 | | 1 | +1 | +1 |
| 7 | 626.77 | | 2326 | | 48 | 627 | | 22 | | | 3 |
| | | 14969 | | 412 | | | 149 | | 7 | -3 | 4 |
| 8 | 776.46 | | 2738 | | 47 | 776 | | 29 | | | |
| | | 17707 | | 459 | | | 178 | | 2 | +3 | 5 |
| 9 | 953.53 | | 3197 | | 46 | 954 | | 31 | | | |
| | | 20904 | | 505 | | | 209 | | 6 | -1 | 5 |
| 10 | 1162.57 | | 3702 | | 45 | 1163 | | 37 | | | |
| | | 24606 | | 550 | | | 246 | | 5 | | 5 |
| 11 | 1408.63 | | 4252 | | | 1409 | | 42 | | | |
| | | 28858 | | | | | 288 | | | | |
| 12 | 1697.21 | | | | | 1697 | | | | | |

This example illustrates that smoothness of differences of rounded values of a function is *not* a guarantee that these values give the best representation of that function. The adjustment of function values by differences cannot be depended on to ± 1 unit in the last place; it is possible, as in this example, to make the differences over-smooth.†

4.45. Direct evaluation of second differences

It is sometimes convenient to be able to evaluate second differences directly from function values without the intermediate step of calculating first differences. This can be done on a machine as follows.

Suppose first that the second differences of f are positive; $\delta^2 f_j$ is calculated from the formula

$$\delta^2 f_j = f_{j-1} + f_{j+1} - 2f_j,$$

the terms being taken in this order; then f_j is set ready for the calculation of

$$\delta^2 f_{j+1} = f_j + f_{j+2} - 2f_{j+1},$$

and so on.

If the second differences are negative, this process will give them in complementary form; then it is more convenient to obtain

$$-\delta^2 f_j = 2f_j - f_{j-1} - f_{j+1},$$

† For a further discussion of checking by differences, see J. C. P. Miller, *M.T.A.C.* **4** (1950), 3.

the terms being taken in this order so that when $-\delta^2 f_j$ has been obtained, f_{j+1} is already set for the calculation of

$$-\delta^2 f_{j+1} = 2f_{j+1} - f_j - f_{j+2}.$$

If the function values are negative, their moduli are set, and the signs of the machine operations altered accordingly.

This is a useful process for checking values of a function built up from second differences by summing the second differences to form the first differences, and then summing first differences to give the function values. The direct calculation of second differences provides a good check of these two successive summations.

4.46. Building up from second differences

A function can be built up directly from its second differences, without calculation of the first differences, by a process which is the converse of that of the previous section. If the function is positive, we have

$$f_{j+1} = 2f_j - f_{j-1} + \delta^2 f_j;$$

this is transferred to the setting levers, and used in the first step in forming

$$f_{j+2} = 2f_{j+1} - f_j + \delta^2 f_{j+1}.$$

If f_j is negative, it is more convenient to form

$$(-f_{j+1}) = 2(-f_j) - (-f_{j-1}) - \delta^2 f_j.$$

If this process of building up a function from its second differences is used, the method of the previous section should *not* be used for checking; the processes are too nearly alike for one to be a good check of the results of the other.

One machine, the Brunsviga 20, has two facilities which are very convenient for building up a function from its second differences; these are transfer from the accumulator to the setting levers, and an arrangement for clearing only the right-hand half of the accumulator, leaving the left-hand half unaffected. The latter feature is known as 'split clearance', and has the effect of furnishing the machine with two registers.

In the present application, the first differences are accumulated in the right-hand half of the accumulator (R.H. for short) and the function itself in the left-hand half (L.H.). Let $\delta f_{j-\frac{1}{2}}$ be in R.H. and f_j in L.H.; then $\delta^2 f_j$ is set on the setting levers (S.L.), and added into R.H., which then contains $\delta f_{j+\frac{1}{2}}$. This is transferred to S.L., and the operation of clearing it from R.H. does not affect L.H.; it is added back into R.H. and also, after shifting the accumulator, into L.H.; R.H. now contains $\delta f_{j+\frac{1}{2}}$ and L.H. contains f_{j+1} . The accumulator is now shifted back, $\delta^2 f_{j+1}$ set, and the

process repeated. The only quantities needing setting are the second differences $\delta^2 f_j$.

4.5. Differences and derivatives

We have seen that functions defined by analytical formulae are adequately represented by tables only in ranges away from singularities and discontinuities, and that if a table is the only information we have about a function, we may regard the function represented by the table as being differentiable as many times as we require. We will therefore suppose that in any application of numerical methods to functions specified by a table, the function can be expanded in a Taylor series over the range with which we are concerned.

Then we have

$$f_{\pm 1} = f_0 \pm (\delta x) f'_0 + \frac{1}{2!} (\delta x)^2 f''_0 \pm \frac{1}{3!} (\delta x)^3 f'''_0 + \dots \quad (4.15)$$

and in general

$$f_{\pm n} = f_0 \pm (n \delta x) f'_0 + \frac{1}{2!} (n \delta x)^2 f''_0 \pm \frac{1}{3!} (n \delta x)^3 f'''_0 + \dots, \quad (4.16)$$

the remainder term being of order $(\delta x)^m$ if the series is cut off after m terms. We shall only derive a few relations directly from these expansions, as we shall shortly see a quicker and more effective way of deriving relations of the kind we require in practical work.

Substitution of series such as (4.15), (4.16) into the formulae giving differences in terms of function values gives a set of relations for *differences in terms of derivatives*; for example, if terms of order $(\delta x)^6$ are included,

$$\begin{aligned} \delta^2 f_0 &= f_1 - 2f_0 + f_{-1} = 2 \left[(\delta x)^2 \frac{1}{2!} f''_0 + \frac{1}{4!} (\delta x)^4 f^{iv}_0 + \frac{1}{6!} (\delta x)^6 f^{vi}_0 \right] + O(\delta x)^8 \\ &= (\delta x)^2 \left[f''_0 + \frac{1}{12} (\delta x)^2 f^{iv}_0 + \frac{1}{360} (\delta x)^4 f^{vi}_0 \right] + O(\delta x)^8, \end{aligned} \quad (4.17)$$

and similarly

$$\delta^4 f_0 = f_2 - 4f_1 + 6f_0 - 4f_{-1} + f_{-2} = (\delta x)^4 \left[f^{iv}_0 + \frac{1}{6} (\delta x)^2 f^{vi}_0 \right] + O(\delta x)^8. \quad (4.18)$$

From the symmetry of the coefficients in formula (4.17) it follows that even-order differences $\delta^{2n} f_0$ involve only even-order derivatives at x_0 .

It follows from (4.17), (4.18) that

$$\begin{aligned} \lim_{\delta x \rightarrow 0} [\delta^2 f_0 / (\delta x)^2] &= f''_0, \\ \lim_{\delta x \rightarrow 0} [\delta^4 f_0 / (\delta x)^4] &= f^{iv}_0, \end{aligned}$$

and similarly for higher orders of differences; thus finite difference *ratios* are closely allied to derivatives. But in using differences, the

differences themselves are the quantities that enter into most formulae, rather than difference ratios.

The relations (4.17), (4.18) and similar ones for higher-order differences can be regarded as equations for derivatives in terms of differences, and solved for these. A more important relation, however, is one between the second differences of f and its second derivative and its differences.

From (4.17) applied to the function f'' we have

$$\delta^2 f''_0 = (\delta x)^2 [f''_0 + \frac{1}{12}(\delta x)^2 f''_0] + O(\delta x)^6,$$

$$\delta^4 f''_0 = (\delta x)^4 f''_0 + O(\delta x)^6,$$

so that

$$(\delta x)^2 f''_0 = \delta^2 f''_0 - \frac{1}{12} \delta^4 f''_0 + O(\delta x)^6,$$

$$(\delta x)^4 f''_0 = \delta^4 f''_0 + O(\delta x)^6,$$

and substitution into (4.17) gives

$$\delta^2 f_0 = (\delta x)^2 [f''_0 + \frac{1}{12} \delta^2 f''_0 - \frac{1}{240} \delta^4 f''_0] + O(\delta x)^8. \quad (4.19)$$

Similarly,

$$\mu \delta f_0 = \frac{1}{2}(f_1 - f_{-1}) = (\delta x) \left[f'_0 + \frac{1}{3!} (\delta x)^2 f'''_0 + \frac{1}{5!} (\delta x)^4 f^{(5)}_0 \right] + O(\delta x)^7, \quad (4.20)$$

and application of (4.17), (4.18) to f' gives

$$\delta^2 f'_0 = (\delta x)^2 [f'_0 + \frac{1}{12} (\delta x)^2 f'_0] + O(\delta x)^6,$$

$$\delta^4 f'_0 = (\delta x)^4 f'_0 + O(\delta x)^6;$$

and substitution in (4.20) gives

$$\mu \delta f_0 = (\delta x) [f'_0 + \frac{1}{6} \delta^2 f'_0 - \frac{1}{180} \delta^4 f'_0] + O(\delta x)^7. \quad (4.21)$$

As we shall see later (§ 6.3), the first two terms in the square bracket here give the formula usually known as 'Simpson's rule' for numerical quadrature.

For relations involving *even-order* differences, and *odd-order mean* differences, the expansions (4.15), (4.16) in f and its derivatives at $x = x_0$ are the most convenient. For corresponding relations involving *odd-order* differences and *even-order mean* differences, it is often more convenient to expand in terms of f and its derivatives at $x = x_{\frac{1}{2}}$.

4.6. Finite difference operators

A powerful method of obtaining formulae for interpolation, integration, etc., in terms of finite differences is by means of finite difference operators. We have already recognized that the symbol δ or Δ prefixed to a symbol representing a function can be regarded as representing an *operation* performed on that function. We will now extend this idea, and first define some further operators.

The operator E is defined by

$$Ef(x) = f(x + \delta x),$$

or shortly

$$Ef_j = f_{j+1}. \quad (4.22)$$

This operator advances the argument from one value to the next of the finite difference table, and is sometimes called the 'shift operator' or 'forward shift operator'. Its inverse, written E^{-1} or $1/E$, the 'backward shift operator', steps the argument back from one value to the previous one in the difference table; that is

$$E^{-1}f_j = f_{j-1}. \quad (4.23)$$

If D is the differential operator $D \equiv d/dx$, Taylor's expansion can be written symbolically

$$f_1 = f(x_0 + \delta x) = e^{(\delta x)D}f_0,$$

so that, formally,

$$Ef_0 = e^{(\delta x)D}f_0 \quad (4.24)$$

for all functions f for which the right-hand side is significant. A relation such as this, between results of different operations, which is independent of the function f operated on, is often written as a relation between the *operators*, without an operand explicitly indicated. We follow this usage, and, in accordance with it write (4.24) as

$$E = e^{(\delta x)D}. \quad (4.25)$$

Two operators of which we shall make considerable use are $E^{\frac{1}{2}}$ and its inverse $E^{-\frac{1}{2}}$. $E^{\frac{1}{2}}$ is the operator which, applied twice to f_0 , gives f_1 , independently of the particular form of the function f ; that is to say, it is an operator such that for any operand f ,

$$E^{\frac{1}{2}}[E^{\frac{1}{2}}f_0] = f_1 = Ef_0.$$

It is clear that an operator which advances the argument value by *half* the tabular interval satisfies this condition; that is

$$E^{\frac{1}{2}}f(x) = f(x + \tfrac{1}{2}\delta x),$$

or

$$E^{\frac{1}{2}}f_0 = f_{\frac{1}{2}}. \quad (4.26)$$

From Taylor's series

$$f_{\frac{1}{2}} = e^{\frac{1}{2}(\delta x)D}f_0$$

so that (4.26) is consistent with (4.25).

The 'forward difference operator' Δ is defined by

$$\Delta f(x) = f(x + \delta x) - f(x) = Ef(x) - f(x),$$

or shortly

$$\Delta f_0 = f_1 - f_0 = (E - 1)f_0$$

which, expressed as a relation between operators, is

$$\Delta = E - 1. \quad (4.27)$$

The 'backward difference operator' ∇ is defined correspondingly by

$$\nabla f(x) = f(x) - f(x - \delta x),$$

or

$$\nabla = 1 - E^{-1} = (E - 1)/E. \quad (4.28)$$

The 'central difference operator' δ is defined by

$$\delta f(x) = f(x + \tfrac{1}{2}\delta x) - f(x - \tfrac{1}{2}\delta x) = (E^{\frac{1}{2}} - E^{-\frac{1}{2}})f(x)$$

which, expressed as a relation between operators, is

$$\delta = E^{\frac{1}{2}} - E^{-\frac{1}{2}}. \quad (4.29)$$

Another useful operator is the 'averaging operator' μ , defined by

$$\mu f(x) = \tfrac{1}{2}[f(x + \tfrac{1}{2}\delta x) + f(x - \tfrac{1}{2}\delta x)] = \tfrac{1}{2}[E^{\frac{1}{2}}f(x) + E^{-\frac{1}{2}}f(x)],$$

i.e.

$$\mu = \tfrac{1}{2}(E^{\frac{1}{2}} + E^{-\frac{1}{2}}). \quad (4.30)$$

These operators all have their inverses. We have already considered the operator inverse to E . The operator inverse to δ is the 'central sum operator' $\sigma = \delta^{-1}$ defined by

$$\sigma f_n = \sigma f_{n-1} + f_{n-\frac{1}{2}},$$

or

$$\sigma = E^{\frac{1}{2}}/(E - 1).$$

It should be noted that σf , like an indefinite integral, is undetermined to the extent of an arbitrary additive constant. The operator inverse to μ will be considered in § 5.2.

These operators are all linear; that is to say if O is any one of them, and f and F are any two functions, then

$$O(f + F) = Of + OF.$$

The operators E , Δ , D , δ , and ∇ are also commutative; that is, if O_1 and O_2 are two of these operations and f is any function,

$$O_1(O_2f) = O_2(O_1f).$$

σ and δ are not necessarily commutative, since $\sigma(\delta f)$ may differ from $\delta(\sigma f)$ by a constant, just as $\int (df/dx) dx$ may differ from f by a constant.

Some useful relations may be obtained from (4.23) to (4.30). For example, from (4.29),

$$\delta^2 = E - 2 + E^{-1}$$

(the operational form of $\delta^2 f_0 = f_1 - 2f_0 + f_{-1}$), and from (4.30)

$$\mu^2 = \tfrac{1}{4}(E + 2 + E^{-1}),$$

whence

$$\delta^2 = 4(\mu^2 - 1),$$

or

$$\mu^2 = 1 + \tfrac{1}{4}\delta^2. \quad (4.31)$$

And if in (4.29), (4.30) we substitute for E from (4.25) we obtain the formal relations

$$\delta = 2 \sinh \frac{1}{2}(\delta x)D, \quad (4.32)$$

$$\mu = \cosh \frac{1}{2}(\delta x)D. \quad (4.33)$$

Also we have

$$\begin{aligned} (E+1)\delta &= E^{\frac{1}{2}}(E^{\frac{1}{2}}+E^{-\frac{1}{2}})(E^{\frac{1}{2}}-E^{-\frac{1}{2}}) \\ &= (E^{\frac{1}{2}}+E^{-\frac{1}{2}})(E-1) = 2(E-1)\mu. \end{aligned} \quad (4.34)$$

Also $\delta^n = [E^{-\frac{1}{2}}(E-1)]^n = E^{-\frac{1}{2}n}(E-1)^n,$

so that $\delta^n f_j = (E-1)^n E^{-\frac{1}{2}n} f_j = (E-1)^n f_{j-\frac{1}{2}n};$

expansion of $(E-1)^n$ by the binomial theorem gives

$$\delta^n f_j = \sum_{k=0}^n (-1)^k \frac{n!}{k!(n-k)!} E^{n-k} f_{j-\frac{1}{2}n} = \sum_{k=0}^n (-1)^k \frac{n!}{k!(n-k)!} f_{j+\frac{1}{2}n-k}$$

in agreement with (4.8).

We shall make considerable use of relations between operators, such as (4.32) and (4.33), which imply the use of Taylor's series in the form

$$e^{\xi D} f(x) = f(x+\xi) \quad (4.35)$$

without a remainder term. However, in using the formulae we finally obtain by means of these relations, we shall in almost all cases retain only the first few terms, thereby making truncation errors in which the remainder term of the Taylor expansion can be considered as incorporated. In most cases an analysis of the truncation error and its relation to the remainder term in the Taylor expansion can be carried out by the method of § 6.8. But it will be as well to know for what kinds of functions this expansion can be used in the form (4.35).

1. *Polynomials*: it is clearly exact for polynomials since the series terminates.
2. *Exponentials*: if $f(x) = e^{ax}$, then

$$e^{\xi D} e^{ax} = \left(1 + \xi D + \frac{1}{2!} \xi^2 D^2 + \dots\right) e^{ax} = \left(1 + \xi a + \frac{1}{2!} \xi^2 a^2 + \dots\right) e^{ax}.$$

The series in the bracket converges for all values of ξ and a , and its value is $e^{a\xi}$. Hence

$$e^{\xi D} f(x) = e^{\xi D} e^{ax} = e^{a\xi} e^{ax} = e^{a(x+\xi)} = f(x+\xi),$$

so that we can apply (4.35) without restriction to exponentials in which the exponent is linear in x .

3. *Products of exponentials and polynomials*: we will prove that if

$$e^{\xi D} z(x) = z(x+\xi),$$

then

$$e^{\xi D} [x z(x)] = (x+\xi) z(x+\xi),$$

so that if (4.35) can be applied to a function $z(x)$, then it can be applied to $x z(x)$; and so, by repetition of the argument, it can be applied to $z(x)$ multiplied by any polynomial. We have

$$\begin{aligned} e^{\xi D} [x z(x)] &\equiv \left[1 + \xi D + \frac{\xi^2}{2!} D^2 + \dots\right] x z(x) \\ &= xz + \xi[xDz + z] + \frac{1}{2!} \xi^2 [xD^2z + 2Dz] + \frac{1}{3!} \xi^3 [xD^3z + 3D^2z] + \dots \\ &= x \left[z + \xi Dz + \frac{1}{2!} \xi^2 D^2z + \frac{1}{3!} \xi^3 D^3z + \dots \right] + \left[\xi z + \xi^2 Dz + \frac{1}{2!} \xi^3 D^2z + \dots \right]. \end{aligned}$$

Since z is assumed to be such that (4.35) is valid, the first square bracket is $xz(x+\xi)$, and the second is $\xi z(x+\xi)$. Hence altogether

$$e^{\epsilon D}[xz(x)] = (x+\xi)z(x+\xi).$$

Thus (4.35) can be applied to products of exponentials (including circular functions) and polynomials; and, since the operator is linear, it can be extended to sums of products of exponentials and polynomials.

4.7. Examples of the use of finite difference operators

It is convenient, for brevity, to have a single symbol for the operator $(\delta x)D$; this will be written U , that is

$$U = (\delta x)D. \quad (4.36)$$

Then the relations (4.25), (4.32), and (4.33) are

$$E = e^U, \quad \delta = 2 \sinh \frac{1}{2}U, \quad \mu = \cosh \frac{1}{2}U, \quad (4.37)$$

$$\text{so that} \quad U = 2 \sinh^{-1} \frac{1}{2}\delta \quad (4.38)$$

$$= [(\sinh^{-1} \frac{1}{2}\delta)/\frac{1}{2}\delta]\delta. \quad (4.39)$$

Since $\delta f = (\delta x)Df + O(\delta x)^3$ for any particular function f to which these relations between finite difference operators can be applied, it follows that in expanding these and other relations in powers of δ or U , δ^n or U^n can be regarded as a quantity of order $(\delta x)^n$.

4.71. Derivatives in terms of differences

Taking the n th power of both sides of (4.39) we have

$$U^n = [(\sinh^{-1} \frac{1}{2}\delta)/\frac{1}{2}\delta]^n \delta^n. \quad (4.40)$$

Since $(\sinh^{-1} z)/z$ is an even function of z , this expresses $U^n = (\delta x)^n D^n$ in *even* powers of δ if n is even, and in *odd* powers of δ if n is odd. The available central differences $\delta^n f_j$ of *even* order have *integral* values of j , whereas those of odd order have $(\text{integer} + \frac{1}{2})$ values of j . Hence this relation can be used to obtain expressions for *even*-order derivatives at tabular values, $D^{2m}f_j$, or *odd*-order derivatives *half-way between* tabular values.

An alternative form is

$$U^n = [\mu^{-1}\{(\sinh^{-1} \frac{1}{2}\delta)/\frac{1}{2}\delta\}^n] \mu \delta^n. \quad (4.41)$$

Since the relation between μ and δ is $\mu^2 = 1 + \frac{1}{4}\delta^2$, the operator in the square bracket is still an even function of δ , so that for *odd* values of n this expresses $U^n f$ in terms of *odd*-order *mean* differences $\mu \delta^{2m+1} f$, which are available at tabular values; hence this is the useful formula for *odd*-order derivatives at tabular values.

The expansions of (4.40) and (4.41) both for positive and for negative values of n can be carried out by taking the series for $(\sinh^{-1} \frac{1}{2}\delta)/\frac{1}{2}\delta$:

$$(\sinh^{-1} \frac{1}{2}\delta)/\frac{1}{2}\delta = 1 - \frac{1}{24}\delta^2 + \frac{3}{640}\delta^4 - \frac{5}{7168}\delta^6 + \frac{35}{264912}\delta^8 + O(\delta x)^{10}$$

and raising it to the appropriate power; and in the case of (4.41) multiplying also by the expansion of $\mu^{-1} = (1 + \frac{1}{4}\delta^2)^{-\frac{1}{2}}$. General expansions for $(U/\delta)^n$ and $[(U/\delta)^n/\mu]$ as far as δ^{10} for any value of n have been given by Bickley;† taken to terms in δ^8 that for $(U/\delta)^n$ is

$$\begin{aligned} \left(\frac{U}{\delta}\right)^n &= \left(\frac{\sinh^{-1} \frac{1}{2}\delta}{\frac{1}{2}\delta}\right)^n \\ &= 1 - \frac{n}{24}\delta^2 + \frac{5n^2 + 22n}{5760}\delta^4 - \frac{35n^3 + 462n^2 + 1528n}{2903040}\delta^6 + \\ &\quad + \frac{175n^4 + 4620n^3 + 40724n^2 + 119856n}{1393459200}\delta^8 + O(\delta x)^{10}. \end{aligned} \quad (4.42)$$

For positive n , the three cases of this formula which we shall need here are the cases $n = 2, 4$, and 6 , namely,

$$(U/\delta)^2 = [1 - \frac{1}{12}\delta^2 + \frac{1}{90}\delta^4 - \frac{1}{560}\delta^6] + O(\delta x)^8, \quad (4.43)$$

$$(U/\delta)^4 = [1 - \frac{1}{6}\delta^2 + \frac{7}{240}\delta^4] + O(\delta x)^6, \quad (4.44)$$

$$(U/\delta)^6 = [1 - \frac{1}{4}\delta^2] + O(\delta x)^4. \quad (4.45)$$

For odd positive powers of (U/δ) , the only important case is $n = 1$, for which

$$(U/\mu\delta) = [1 - \frac{1}{6}\delta^2 + \frac{1}{30}\delta^4 - \frac{1}{140}\delta^6 + \frac{1}{830}\delta^8] + O(\delta x)^{10}. \quad (4.46)$$

These formulae, which give powers of $U = (\delta x)D$ in terms of δ , are operational forms of formulae for differentiation, since, applied to a function f , they give $D^n f$ in terms of the differences of f . They have, however, other and more important applications as will be seen in the next section and subsequent chapters.

4.72. Negative powers of (U/δ)

Other important relations are some involving negative powers of (U/δ) . One way of obtaining these is by use of formula (4.42), or Bickley's corresponding formula for $(U/\delta)^n/\mu$, with negative values of n , for which these formulae are also valid. For example, substitution of $n = -2$ in (4.42) gives

$$(U/\delta)^{-2} = [1 + \frac{1}{12}\delta^2 - \frac{1}{240}\delta^4 + \frac{31}{60480}\delta^6 - \frac{289}{3628800}\delta^8] + O(\delta x)^{10}. \quad (4.47)$$

Another procedure is first to express the operator in terms of U , expand as a series in U , and then substitute in terms of δ from formulae (4.43) to (4.46); this only involves the use of formula (4.42) for positive values of n .

† W. G. Bickley, *Journ. Math. and Phys.* **27** (1948), 183.

The main operators which involve inverse powers of U and for which we require expressions in terms of δ are $(\delta/U)^2$, $(\delta/\mu U)$, and $(\mu\delta/U)$. In terms of U they are

$$(\delta/U)^2 = [(\sinh \tfrac{1}{2}U)/\tfrac{1}{2}U]^2 = 2(\cosh U - 1)/U^2, \quad (4.48)$$

$$(\delta/\mu U) = (\tanh \tfrac{1}{2}U)/\tfrac{1}{2}U, \quad (4.49)$$

$$(\mu\delta/U) = (\sinh U)/U. \quad (4.50)$$

Expansion of (4.48) in powers of U gives

$$\begin{aligned} (\delta/U)^2 &= 1 + \frac{2}{4!}U^2 + \frac{2}{6!}U^4 + \frac{2}{8!}U^6 + O(\delta x)^8 \\ &= 1 + \frac{1}{12}\delta^2(U/\delta)^2 + \frac{1}{360}\delta^4(U/\delta)^4 + \frac{1}{20160}\delta^6(U/\delta)^6 + O(\delta x)^8. \end{aligned}$$

Substitution from formulae (4.43) to (4.45) then gives, to terms in δ^6 ,

$$\begin{aligned} (\delta/U)^2 &= 1 + \frac{1}{12}\delta^2[1 - \frac{1}{12}\delta^2 + \frac{1}{90}\delta^4] + \frac{1}{360}\delta^4[1 - \frac{1}{6}\delta^2] + \frac{1}{20160}\delta^6 + O(\delta x)^8 \\ &= [1 + \frac{1}{12}\delta^2 - \frac{1}{240}\delta^4 + \frac{31}{60480}\delta^6] + O(\delta x)^8 \end{aligned}$$

in agreement with (4.47).

Similarly the following expansions can be obtained:

$$(\delta/\mu U) = [1 - \frac{1}{12}\delta^2 + \frac{11}{720}\delta^4 - \frac{191}{80480}\delta^6 + \frac{2497}{3628800}\delta^8] + O(\delta x)^{10}, \quad (4.51)$$

$$(\mu\delta/U) = [1 + \frac{1}{6}\delta^2 - \frac{1}{180}\delta^4 + \frac{1}{1512}\delta^6 - \frac{23}{228800}\delta^8] + O(\delta x)^{10}. \quad (4.52)$$

The latter of these can alternatively be obtained from the former by multiplying by $\mu^2 = 1 + \frac{1}{4}\delta^2$.

4.73. $\delta^2 f$ in terms of f'' and its differences

We will use some of these relations between differential operators and finite difference operators to express $\delta^2 f_0$ in terms of f''_0 and the central differences of f'' at $x = x_0$. A first approximation is

$$\delta^2 f_0 = (\delta x)^2 f''_0 = (\delta x)^2 D^2 f_0 = U^2 f_0;$$

to improve on this we must find an operator $\phi(\delta)$ such that

$$\delta^2 f_0 = \phi(\delta) U^2 f_0. \quad (4.53)$$

The operator $\phi(\delta)$ required is therefore

$$\phi(\delta) = (\delta/U)^2;$$

its expansion in powers of δ is given by (4.47) above, and substitution in (4.53) gives

$$\delta^2 f_0 = (\delta x)^2 [f''_0 + \frac{1}{12}\delta^2 f''_0 - \frac{1}{240}\delta^4 f''_0 + \frac{31}{60480}\delta^6 f''_0 - \frac{289}{3628800}\delta^8 f''_0] + O(\delta x)^{12} \quad (4.54)$$

(compare formula (4.19) and its derivation in § 4.5).

4.74. $\delta f_{\frac{1}{2}}$ symmetrically in terms of f' and its differences at x_0 and x_1

By definition, $\delta f_{\frac{1}{2}} = f_1 - f_0 = (E-1)f_0$, and to a first approximation

$$\delta f_{\frac{1}{2}} = \frac{1}{2}(\delta x)(f'_0 + f'_1) = \frac{1}{2}(\delta x)(E+1)Df_0 = \frac{1}{2}(E+1)Uf_0; \quad (4.55)$$

we want to obtain a more general relation of which this is the leading term.

By a formula symmetrical in f' and its derivatives at the two ends of the tabular interval is meant one in which the coefficient a_n of each $\delta^n f'_1$ is the same as that of the corresponding $\delta^n f'_0$, so that these terms together give a contribution $a_n(\delta^n f'_0 + \delta^n f'_1) = a_n \delta^n (E+1)Df_0$, as the terms with $n = 0$ do in the first approximation (4.55). Hence we want a relation of the form

$$\delta f_{\frac{1}{2}} \equiv (E-1)f_0 = \frac{1}{2}\phi(\delta)(E+1)Uf_0; \quad (4.56)$$

to satisfy (4.56), $\phi(\delta)$ must be given by

$$\phi(\delta) = \frac{E-1}{\frac{1}{2}(E+1)U}. \quad (4.57)$$

This can be expressed in terms of U by substituting $E = e^U$; this gives

$$\phi(\delta) = (\tanh \frac{1}{2}U)/\frac{1}{2}U.$$

Alternatively, it follows from (4.34) that (4.57) can be written

$$\phi(\delta) = \delta/\mu U$$

for which the expansion in powers of δ is given by (4.51). Hence the required formula is

$$f_1 - f_0 = \frac{1}{2}(\delta x)[f'_0 + f'_1 - \frac{1}{12}(\delta^2 f'_0 + \delta^2 f'_1) + \frac{11}{720}(\delta^4 f'_0 + \delta^4 f'_1) - \frac{191}{60480}(\delta^6 f'_0 + \delta^6 f'_1)] + O(\delta x)^9. \quad (4.58)$$

This is an integration formula, for if $f'(x)$ is given as a function of x it enables the change in $f(x)$, that is $\int_{x_0}^{x_1} f'(x) dx$, to be evaluated in terms of the values of $f'(x)$ (see Chapter VI).

4.75. $\mu \delta f_0$ in terms of f' and its differences at $x = x_0$

In this case we want to find an operator $\phi(\delta)$ such that

$$\mu \delta f_0 = (\delta x)\phi(\delta)Df_0.$$

The appropriate $\phi(\delta)$ is

$$\phi(\delta) = \mu \delta / U = (\sinh U)/U$$

of which the expansion in powers of δ is given by (4.52). Hence

$$\mu\delta f_0 = \frac{1}{2}(f_1 - f_{-1}) = (\delta x)\left[f'_0 + \frac{1}{6}\delta^2 f'_0 - \frac{1}{180}\delta^4 f'_0 + \frac{1}{1512}\delta^6 f'_0 - \frac{23}{226800}\delta^8 f'_0\right] + O(\delta x)^{11} \quad (4.59)$$

(compare § 4.5, formula (4.21)). This also is an integration formula, relating the change of f in an interval $2\delta x$ of x to the behaviour of its derivative in the neighbourhood of that interval.

INTERPOLATION

5.1. Linear and non-linear interpolation

GIVEN a table of values of a function $f(x)$ at a set of tabular values of x , usually, but not necessarily, equally spaced, we may require to determine either the value of $f(x)$ at an intermediate value of x , or the value of x for which $f(x)$ has some specified value. The process for finding a result of this kind is called 'interpolation', and, when it is necessary to distinguish between them, the former is called 'direct' and the latter 'inverse' interpolation. The distinction is not usually significant unless the tabular values of x are equally spaced; this case, however, is much the most usual.

By 'linear interpolation' is meant interpolation using the approximation in which, for $0 < p < 1$, we take

$$f(x_0 + p\delta x) = f_0 + p\delta f_1; \quad (5.1)$$

expressed graphically, this is interpolation along the chord joining the points (x_0, f_0) and (x_1, f_1) . This process is valid so long as the tabular values of x are spaced closely enough; we will obtain later (§ 5.22) a quantitative criterion of what is 'closely enough' in this context. 'Non-linear interpolation' is interpolation in some form which takes account of the departure of the (x, f) curve from the chord between the points corresponding to neighbouring tabular values.

There are two kinds of tables; first, those in which interpolation is required frequently enough to justify the use of intervals of the argument small enough for linear interpolation to be adequate; secondly, those in which interpolation will only be occasional, not frequent enough to justify the calculation and printing at small enough intervals for linear interpolation to be applicable. In the latter case, non-linear interpolation is necessary. But if non-linear interpolation were generally recognized as a standard process, the bulk of tables could be very greatly reduced. For example, a table of $\sin x$ to five decimals at intervals of 10° reads as given on p. 60. We shall see later that the formulae required for carrying out non-linear interpolation in this table are comparatively simple. We shall also see that for linear interpolation we require $|\delta^2 f|$ to be not greater than 2, so that at least 40 times the number of entries are required in order to obtain a table in which linear interpolation can be carried out.

| x | $f(x) = \sin x$ | δf | $\delta^2 f$ | $\delta^3 f$ | $\delta^4 f$ |
|------------|-----------------|------------|--------------|--------------|--------------|
| 0° | 0 | | 0 | | 0 |
| | | 17365 | | -528 | |
| 10° | .17365 | | -528 | | +17 |
| | | 16837 | | -511 | |
| 20° | .34202 | | -1039 | | 31 |
| | | 15798 | | -480 | |
| 30° | .50000 | | -1519 | | 45 |
| | | 14279 | | -435 | |
| 40° | .64279 | | -1954 | | 63 |
| | | 12325 | | -372 | |
| 50° | .76604 | | -2326 | | 65 |
| | | 9999 | | -307 | |
| 60° | .86603 | | -2633 | | 86 |
| | | 7366 | | -221 | |
| 70° | .93969 | | -2854 | | 82 |
| | | 4512 | | -139 | |
| 80° | .98481 | | -2993 | | 94 |
| | | 1519 | | -45 | |
| 90° | 1.00000 | | -3038 | | 90 |

The reduction in bulk achieved by the use of a large interval and non-linear interpolation is not important in the case of functions of a single real variable, but becomes important in connexion with functions of two variables (or of a complex variable), or functions of a variable and one or more parameters such as the Bessel functions $J_n(x)$ and the Whittaker functions $W_{k,m}(x)$.

5.11. Linear interpolation

The simplest form of interpolation is linear interpolation, or interpolation by proportional parts, for which the interpolation formula is (5.1) above.

In carrying out linear interpolation on a machine, there is a precaution against mistakes which should always be observed. Suppose first that $\delta f_{\frac{1}{2}}$ is positive. Having cleared the accumulator, set f_0 , add it into the accumulator, and clear the multiplier register. Then set $\delta f_{\frac{1}{2}}$, *add it in, and verify that the content of the accumulator is now f_1* . This checks that the right values of f_0 and $\delta f_{\frac{1}{2}}$ have been taken. If p has m decimals, the accumulator should first be shifted m places right, and f_0 and $\delta f_{\frac{1}{2}}$ then set on the extreme right of the setting levers or keyboard.

For direct interpolation, $p \delta f_{\frac{1}{2}}$ is added to f_0 ; if p is greater than $\frac{1}{2}$, the addition of $\delta f_{\frac{1}{2}}$ to f_0 to check can be taken as the first step of this multiplication; if p is less than $\frac{1}{2}$, $\delta f_{\frac{1}{2}}$ should be subtracted from f_1 to restore f_0 before doing the multiplication.

For inverse interpolation, the given value of f is built up in the accumulator, and the fraction p of the interval length required to give this value of f is read on the multiplier register.

If $\delta f_{\frac{1}{2}}$ is negative, $|\delta f_{\frac{1}{2}}|$ should be set, and operations of addition and subtraction are interchanged; otherwise the procedure is the same.

In some tables, particularly elementary ones, a sequence of function values is given on a single line (for example, $\log 1.00$ to 1.09 on a line of a four-figure table of logarithms) with proportional parts of the *mean* first difference at the end of the line. Use of these proportional parts of the mean difference does not usually give the best interpolated value, and should not be used indiscriminately except in contexts in which an error of 2 or 3 units in the last figure is unimportant. The following example is taken from a table of logarithms to five places in which, for $x = 1.0$ to 2.0 , different sets of proportional parts of mean differences are given for every *five* entries:

| | | | | | | | | | | | | | |
|----------|--------|--------|--------|--------|--------|--|----|----|-----|-----|-----|-----|-----|
| x | 1.05 | 1.06 | 1.07 | 1.08 | 1.09 | | 1 | 2 | 3 | ... | 5 | ... | 9 |
| $\log x$ | .02119 | .02531 | .02938 | .03342 | .03743 | | 40 | 81 | 121 | ... | 202 | ... | 364 |

The last five columns are the proportional parts of the mean difference. Using the actual difference between the first two entries, we get $\log 1.055 = 0.02325$, whereas using the proportional parts of mean differences we get 0.02321 , a difference of four units in the fifth decimal.

For the best linear interpolation, proportional parts should be taken of the actual difference between successive tabular values. Tables of proportional parts for this purpose are given in most good modern books of tables.

If several functions are tabulated in parallel columns, at such an interval that linear interpolation can be used on each of them, then linear interpolation can be used between two columns. For example, a table of $\sin x$ and $\cos x$ against x , in parallel columns, is also a table of $(1-y^2)^{\frac{1}{2}}$ against y , and can be used as such without reference to the x column at all. Since the values of both functions $f(x)$ and $g(x)$ are subject to rounding error, the possible error in the interpolated value is rather greater than if $f(x)$ were tabulated at exact values of $g(x)$.

5.2. Non-linear interpolation

In considering non-linear interpolation, it will be supposed for the present that the tabular values of the argument are equally spaced.† Interpolation with unequally spaced values of the argument will be considered in § 5.7.

5.21. Half-way interpolation

One particular case of non-linear interpolation is so much simpler than the general case, and so useful, that it will be considered separately

† For a fuller treatment of non-linear interpolation, see L. Fox, *The Use and Construction of Mathematical Tables* (H.M.S.O., 1956).

first. This is interpolation for a value of x half-way between tabular values.

To get a formula for this, we want to express $f(x_0 + \frac{1}{2}\delta x)$, which can be expressed as $E^{\frac{1}{2}}f_0$, symmetrically in f_0, f_1 and the differences of f at x_0 and x_1 . Now $f(x_0 + \frac{1}{2}\delta x) = E^{\frac{1}{2}}f_0$, so we want to find an operator $\phi(\delta)$ such that

$$E^{\frac{1}{2}}f_0 = \phi(\delta)\frac{1}{2}(1+E)f_0.$$

This operator is given by

$$\text{so that } \phi(\delta) = 2E^{\frac{1}{2}}/(1+E) = 1/\cosh \frac{1}{2}U = 1/(1+\frac{1}{4}\delta^2)^{\frac{1}{2}}, \quad (5.2)$$

$$\begin{aligned} f_{\frac{1}{2}} &= f(x_0 + \frac{1}{2}\delta x) = (1 + \frac{1}{4}\delta^2)^{-\frac{1}{2}}[\frac{1}{2}(f_0 + f_1)] \\ &= \frac{1}{2}[f_0 + f_1 - \frac{1}{8}(\delta^2 f_0 + \delta^2 f_1) + \frac{3}{128}(\delta^4 f_0 + \delta^4 f_1) - \\ &\quad - \frac{5}{1024}(\delta^6 f_0 + \delta^6 f_1)] + O(\delta x)^8. \end{aligned} \quad (5.3)$$

It will be noted that the operator $\phi(\delta)$ given by (5.2) is the operator inverse to the averaging operator μ . Indeed, the relation (5.3) could be obtained as follows. The definition of the operator μ is $\mu f_{\frac{1}{2}} = \frac{1}{2}(f_0 + f_1)$, and it follows that the inverse operator μ^{-1} is an operator such that

$$f_{\frac{1}{2}} = \frac{1}{2}\mu^{-1}(f_0 + f_1). \quad (5.4)$$

But $f_{\frac{1}{2}} = f(x_0 + \frac{1}{2}\delta x)$ and $\mu^2 = 1 + \frac{1}{4}\delta^2$, so that (5.4) is just (5.2) in a different form.

Formula (5.3), perhaps taken to higher orders of differences, is useful in a preliminary breaking down of the interval of a table of a function evaluated at a large interval, before carrying out a subtabulation. The coefficients are easy to calculate and to check if more are required than are given in (5.3).† If $(-)^j a_j$ is the coefficient of $(\delta^{2j}f_0 + \delta^{2j}f_1)$ in the square bracket in (5.3) then

$$a_{j+1}/a_j = (2j+1)/8(j+1) \quad (5.5)$$

and the coefficients can most conveniently be calculated by continued multiplication by the successive ratios (5.5). A check is given by the relations $\sum a_j = 2/\sqrt{3} = 1.15470$. $\sum 2^j a_j = \sqrt{2} = 1.41421$.

It is interesting to examine the result of applying (5.3) to a table of $\cos x$ or $\sin x$ at a large interval such as 60° or 90° . The ratio (5.5) tends to the value $\frac{1}{4}$ for large j . Hence provided $|\delta^{2j+2}f/\delta^{2j}f| < 4$ for large j , the infinite series of which (5.3) gives the first few terms formally converges.

Now if $f(x) = B \cos(x + \beta)$, then $\delta^2 f_j = -2(1 - \cos \delta x)f_j$, so that if $\delta x = \frac{1}{2}\pi$, $|\delta^{2j+2}f/\delta^{2j}f| = 2$. Thus by use of the series (5.3) we can interpolate $\cos x$ and $\sin x$, not only approximately but, by taking enough terms, to any accuracy we require, from the tabular values

| | | | | | |
|----------|---|------------------|-------|------------------|--------|
| x | 0 | $\frac{1}{2}\pi$ | π | $\frac{3}{2}\pi$ | 2π |
| $\cos x$ | 1 | 0 | -1 | 0 | 1 |
| $\sin x$ | 0 | 1 | 0 | -1 | 0 |

and the condition of periodicity.

† See also *Interpolation and Allied Tables* (1956 edition) § C6, p. 58.

It is even possible to interpolate accurately from a table at intervals of $\frac{2}{3}\pi$:

| | | | | |
|----------|---|-----------------------|------------------------|--------|
| x | 0 | $\frac{2}{3}\pi$ | $\frac{4}{3}\pi$ | 2π |
| $\cos x$ | 1 | $-\frac{1}{2}$ | $-\frac{1}{2}$ | 1 |
| $\sin x$ | 0 | $\frac{1}{2}\sqrt{3}$ | $-\frac{1}{2}\sqrt{3}$ | 0 |

extended by using the condition of periodicity.

5.22. Newton's forward-difference formula

Of the formulae for non-linear interpolation for a general value of the fraction p of the interval length, the simplest to derive is one in terms of forward differences. Its practical value is, however, limited.

Taylor's series can be written, in terms of operators,

$$f(x_0 + p\delta x) = e^{p(\delta x)D}f_0 = E^p f_0.$$

Also $E = 1 + \Delta$, and expansion of $(1 + \Delta)^p$ by the binomial theorem gives

$$\begin{aligned} f(x_0 + p\delta x) &= \left[1 + p\Delta + \frac{1}{2!}p(p-1)\Delta^2 + \dots \right] f_0 \\ &= f_0 + p\Delta f_0 + \frac{1}{2!}p(p-1)\Delta^2 f_0 + \frac{1}{3!}p(p-1)(p-2)\Delta^3 f_0 + \dots \quad (5.6) \end{aligned}$$

which is usually known as Newton's formula. It uses values of the differences on an inclined line in a difference table:

| | | | | | |
|-------|-------|--------------|----------------|----------------|----------------|
| x | f | | | | |
| x_0 | f_0 | | | | |
| x_1 | f_1 | Δf_0 | $\Delta^2 f_0$ | $\Delta^3 f_0$ | $\Delta^4 f_0$ |
| x_2 | f_2 | Δf_1 | $\Delta^2 f_1$ | $\Delta^3 f_1$ | $\Delta^4 f_1$ |
| x_3 | f_3 | Δf_2 | $\Delta^2 f_2$ | $\Delta^3 f_2$ | $\Delta^4 f_2$ |

It is unsatisfactory if differences beyond the second have to be taken into account, as the differences of a function f depend primarily on the behaviour of the function in the neighbourhood of the value of x on which they are centred, so that the higher-order differences involved in this formula are less and less closely related to the behaviour of f in the interval in which interpolation is being carried out.

Its practical use is restricted to interpolation near the boundaries of a table, and this is rare because unless f or one of its derivatives has a singularity at $x = x_0$, there should usually be little difficulty in extending the table backwards a few intervals from $x = x_0$, whereas if the boundary of the table results from f being infinite at $x = x_0$ (for example $f(x) = \cot x$ at $x = 0$) or undefined for $x < x_0$ (for example $f(x) = x^{\frac{1}{2}}$ at $x = 0$), this situation is usually associated with an infinite derivative $f'(x_0)$, in which case the Taylor series expansion on which Newton's formula is based is invalid.

There are various other interpolation formulae, which can all be derived from Newton's by substitution for the forward differences $\Delta^n f_0$ in terms of differences more representative of the behaviour of $f(x)$ in the interval through which the interpolation is being carried out. It is difficult, however, to obtain the form of the general term by such a derivation, and it is better to derive these other interpolation formulae independently. Of the various formulae Comrie writes† 'only three are found in good modern practice, namely those associated with the names of Bessel and Everett, each of which is a simple transformation of the other, and that of Lagrange'. The present treatment will be restricted to these three.

From Newton's (or Bessel's) formula it is possible to deduce the conditions in which linear interpolation gives a sufficiently accurate result. The greatest numerical value of the coefficient of the second difference in formula (5.6) is $\frac{1}{8}$. It is best to keep the contribution from this term to the interpolated value less than 0.3 in the last figure; if it were greater it should be included as it might affect the rounding off of the final result. Hence linear interpolation should not be used if second differences are greater than 2 unless errors up to 2 units in the last place of the interpolated value can be tolerated.

Occasionally the contribution from the second differences to the interpolated value is negligible when those from higher orders of differences are not; an example is provided by the function $x(x^2-1)(x^2-4)$ tabulated at unit intervals of x and interpolated between $x = 0$ and 1. To avoid this situation it is only necessary to see that not only the second differences used in the interpolation formula, but also a number of neighbouring values, are not greater than 2 in the last figure.

5.3. Some expansions

For the purpose of deriving interpolation formulae in central differences,‡ we shall require some expansions, namely those of $\sinh \beta U$, $(\cosh \beta U)/\cosh \frac{1}{2}U$, and $(\sinh \beta U)/\sinh U$ in terms of $\delta = 2 \sinh \frac{1}{2}U$, for non-integral β . These could be written down from the similar expressions for circular functions of a numerical variable;§ but their derivations will be given here for completeness.

For the purposes of this section, let u stand for an ordinary numerical variable, and let

$$z = 2 \sinh \frac{1}{2}u \quad \text{and} \quad y = \cosh \beta u. \quad (5.7)$$

† *Chambers's 6-Figure Tables*, vol. 2 (1949), Introduction, p. xxvii.

‡ The treatment of this and the following section follows that of J. G. L. Michel, *Journ. Inst. of Actuaries*, 72 (1946), 470.

§ e.g. T. J. I'A. Bromwich, *Theory of Infinite Series* (Macmillan, 2nd ed. 1926) § 68.

Consider first the expansion of y as a power series in z . We shall obtain this by forming the differential equation for y in terms of z as independent variable, then differentiating n times and putting $z = 0$; this will give recurrence relations for the derivatives $d^n y/dz^n$ at $z = 0$, from which their values, and so the required series, can be written down.

Since $y = \cosh \beta u$, it satisfies the equation

$$\frac{d^2 y}{du^2} = \beta^2 y, \quad (5.8)$$

and since $z = 2 \sinh \frac{1}{2}u$, it follows that

$$\frac{dz}{du} = \cosh \frac{1}{2}u, \quad (5.9)$$

so that
$$\frac{d^2 y}{du^2} = (\cosh \frac{1}{2}u) \frac{d}{dz} \left[(\cosh \frac{1}{2}u) \frac{dy}{dz} \right].$$

On differentiating this out, substituting for $\sinh \frac{1}{2}u$ from (5.7) and for du/dz from (5.9), we obtain

$$(1 + \frac{1}{4}z^2) \frac{d^2 y}{dz^2} + \frac{1}{2}z \frac{dy}{dz} = \beta^2 y, \quad (5.10)$$

and then, on differentiating n times with respect to z and putting $z = 0$,

$$y^{(n+2)}(0) = (\beta^2 - \frac{1}{4}n^2)y^{(n)}(0). \quad (5.11)$$

Also, for small u , $z = u + O(u^3)$, and so

$$y = 1 + O(u^2) = 1 + O(z^2)$$

and hence $y(0) = 1$, $y'(0) = 0$. Hence, from (5.11), for the odd derivatives

$$y^{(2n+1)}(0) = 0$$

and for the even derivatives

$$y''(0) = \beta^2, \quad y^{(4)}(0) = \beta^2(\beta^2 - 1), \quad y^{(6)}(0) = \beta^2(\beta^2 - 1)(\beta^2 - 4) \dots,$$

and in general, in terms of the functions ψ_m introduced in § 4.41,

$$y^{(2n)}(0) = \beta \psi_{2n-1}(\beta). \quad (5.12)$$

Hence
$$y = \cosh \beta u = 1 + \beta \sum_n \psi_{2n+1}(\beta) z^{2n+2}/(2n+2)!.$$

Differentiation with respect to z then gives

$$\beta (\sinh \beta u) \frac{du}{dz} = \beta z \sum_n \psi_{2n+1}(\beta) z^{2n}/(2n+1)!.$$

But from (5.9)

$$z dz/du = (2 \sinh \frac{1}{2}u) \cosh \frac{1}{2}u = \sinh u;$$

hence

$$\frac{\sinh \beta u}{\sinh u} = \sum_n \psi_{2n+1}(\beta) z^{2n} / (2n+1)! \quad (5.13)$$

$$= \beta \left[1 + \frac{1}{3!} (\beta^2 - 1) z^2 + \frac{1}{5!} (\beta^2 - 1)(\beta^2 - 4) z^4 + \dots \right]. \quad (5.14)$$

To obtain corresponding expressions for $\sinh \beta u$ and $\cosh \beta u / \cosh \frac{1}{2} u$, take $z = 2 \sinh \frac{1}{2} u$ as before, and $y = \sinh \beta u$. This also satisfies equation (5.8), and the above argument applies as far as the recurrence relation (5.11). Now, however, $y = \beta z + O(z^3)$ for small z , so that $y(0) = 0$, $y'(0) = \beta$. Hence $y^{(2n)}(0) = 0$, and

$$y'(0) = \beta, \quad y'''(0) = \beta(\beta^2 - \frac{1}{4}), \quad y^{(5)}(0) = \beta(\beta^2 - \frac{1}{4})(\beta^2 - \frac{9}{4}), \quad \dots,$$

and in general, in terms of the functions ψ_n of § 4.41 (p. 40),

$$y^{(2n+1)}(0) = \beta \psi_{2n}(\beta)$$

(compare (5.12) for the expansion of $\cosh \beta u$). Hence

$$y = \sinh \beta u = \beta \sum_n \psi_{2n}(\beta) z^{2n+1} / (2n+1)! \quad (5.15)$$

$$= \beta \left[z + \frac{1}{3!} \left(\beta^2 - \frac{1}{4} \right) z^3 + \frac{1}{5!} \left(\beta^2 - \frac{1}{4} \right) \left(\beta^2 - \frac{9}{4} \right) z^5 + \dots \right]. \quad (5.16)$$

The expansion of $(\cosh \beta u) / (\cosh \frac{1}{2} u)$ can be obtained by differentiating (5.15) with respect to z . On the left-hand side this gives $\beta (\cosh \beta u) (du/dz)$. But from (5.9) this is just $\beta (\cosh \beta u) / (\cosh \frac{1}{2} u)$. Hence

$$\frac{\cosh \beta u}{\cosh \frac{1}{2} u} = \sum_n \psi_{2n}(\beta) z^{2n} / (2n)! = 1 + \frac{1}{2!} \left(\beta^2 - \frac{1}{4} \right) z^2 + \frac{1}{4!} \left(\beta^2 - \frac{1}{4} \right) \left(\beta^2 - \frac{9}{4} \right) z^4 + \dots \quad (5.17)$$

5.4. Everett's interpolation formula

The simplest central-difference interpolation formula to obtain is that known as Everett's. This expresses the interpolated value of f in terms of the values of f and of its *even*-order differences only, at the beginning and end of the interval in which the interpolation is being carried out; that is, it is of the form

$$f(x_0 + p \delta x) = (1-p)f_0 + pf_1 + E_2(p)\delta^2 f_0 + F_2(p)\delta^2 f_1 + \\ + E_4(p)\delta^4 f_0 + F_4(p)\delta^4 f_1 + \dots \quad (5.18)$$

The coefficients in this interpolation formula† are usually known as 'Everett interpolation coefficients'; they are functions of the fraction p of the interval length δx for which the interpolation is being carried out.

To obtain a formula of this kind we must find operators $\phi_0(\delta)$, $\phi_1(\delta)$ which involve only even powers of δ and which are such that

$$f(x_0 + p \delta x) = \phi_0(\delta)f_0 + \phi_1(\delta)f_1. \quad (5.19)$$

† The notation p for the fraction of the interval δx for which interpolation is carried out, $q = 1 - p$, and E_{2n} and F_{2n} for the coefficients in the Everett formula, here used is adopted to conform to that of *Interpolation and Allied Tables* (H.M. Nautical Almanac Office), 1956, and L. Fox 'The Use and Construction of Mathematical Tables' (H.M.S.O., 1956). The coefficients here written E_{2n} and F_{2n} are also written E_0^{2n} and E_1^{2n} .

Now $f(x_0 + p \delta x) = e^{pU} f_0$, and $f_1 = e^U f_0$, so (5.19), expressed as a relation between operators, becomes

$$e^{pU} = \phi_0(\delta) + \phi_1(\delta)e^U.$$

Since $\phi_0(\delta)$, $\phi_1(\delta)$ are to be even functions of δ , and so of U , it follows that they do not change on replacing U by $-U$; hence

$$e^{-pU} = \phi_0(\delta) + \phi_1(\delta)e^{-U},$$

and solution for $\phi_0(\delta)$, $\phi_1(\delta)$ then gives

$$\phi_1(\delta) = (e^{pU} - e^{-pU}) / (e^U - e^{-U}) = \sinh pU / \sinh U,$$

$$\phi_0(\delta) = \sinh(1-p)U / \sinh U = \sinh qU / \sinh U,$$

where, for convenience later, q has been written for $(1-p)$.

We require expressions for these operators in terms of δ . These could be obtained by expanding in powers of U and then substituting for U in terms of δ from formulae (4.43) to (4.45); but the form of the general term is most easily obtained by the formal substitution of U for u , δ for z , and p or $q = 1-p$ for β in (5.13) or (5.14). This, followed by substitution of the results into (5.19), gives

$$f(x_0 + p \delta x) = \sum_n \frac{1}{(2n+1)!} [\psi_{2n+1}(1-p)\delta^{2n}f_0 + \psi_{2n+1}(p)\delta^{2n}f_1] \quad (5.20)$$

$$\begin{aligned} &= q \left[f_0 + \frac{1}{3!} (q^2-1)\delta^2 f_0 + \frac{1}{5!} (q^2-1)(q^2-4)\delta^4 f_0 + \dots \right] + \\ &+ p \left[f_1 + \frac{1}{3!} (p^2-1)\delta^2 f_1 + \frac{1}{5!} (p^2-1)(p^2-4)\delta^4 f_1 + \dots \right] \end{aligned} \quad (5.21)$$

which is Everett's interpolation formula. The leading terms, $qf_0 + pf_1$, give the value $f_0 + p\delta f_{\frac{1}{2}}$ obtained by linear interpolation, expressed in a form consistent with the other terms of the formula.

Comparison of (5.20) with (5.18) gives the following general expressions for the coefficients in (5.18):

$$\left. \begin{aligned} E_{2n}(p) &= \psi_{2n+1}(1-p)/(2n+1)! \\ F_{2n}(p) &= \psi_{2n+1}(p)/(2n+1)! \end{aligned} \right\}. \quad (5.22)$$

5.41. Bessel's interpolation formula

Bessel's interpolation formula expresses the interpolated value $f(x_0 + p \delta x)$ in terms of mean differences of even-order $\mu\delta^{2n}f_{\frac{1}{2}}$ and odd-order differences $\delta^{2n+1}f_{\frac{1}{2}}$, centred on the middle of the interval in which interpolation is being carried out. For practical work it is most convenient to have the contribution from the even-order differences expressed in

terms of the sum $(\delta^{2n}f_0 + \delta^{2n}f_1)$ of the values at the beginning and end of the interval. Thus this formula is of the general form†

$$f(x_0 + p \delta x) = \frac{1}{2}(f_0 + f_1) + (p - \frac{1}{2})\delta f_{\frac{1}{2}} + B_2(p)(\delta^2 f_0 + \delta^2 f_1) + \\ + B_3(p)\delta^3 f_{\frac{1}{2}} + B_4(p)(\delta^4 f_0 + \delta^4 f_1) + B_5(p)\delta^5 f_{\frac{1}{2}} + \dots \quad (5.23)$$

The coefficients $B_n(p)$ of successive orders of differences in this formula are called ‘Bessel interpolation coefficients’, or simply ‘Bessel coefficients’ when there is no danger of confusion with the other meaning of this term. The first two terms give the value $f_0 + p \delta f_{\frac{1}{2}}$ obtained by linear interpolation, expressed in a form consistent with the other terms of the series.

A formula of this kind can be derived quite easily from Everett’s formula (5.18). Consider the pairs of terms involving $\delta^{2n}f_0$ and $\delta^{2n}f_1$ in Everett’s formula. These can be written

$$E_{2n}(p)\delta^{2n}f_0 + F_{2n}(p)\delta^{2n}f_1 \\ = \frac{1}{2}[E_{2n}(p) + F_{2n}(p)](\delta^{2n}f_0 + \delta^{2n}f_1) + \frac{1}{2}[F_{2n}(p) - E_{2n}(p)]\delta^{2n+1}f_{\frac{1}{2}},$$

which is of the form of the contributions from $\delta^{2n}f$ and $\delta^{2n+1}f$ in Bessel’s formula. Comparison with (5.23) and use of the formulae (5.22) for the Everett coefficients gives

$$B_{2n}(p) = \frac{1}{2}[E_{2n}(p) + F_{2n}(p)] \\ = \frac{1}{2} \frac{1}{(2n+1)!} [\psi_{2n+1}(1-p) + \psi_{2n+1}(p)]$$

and

$$B_{2n+1}(p) = \frac{1}{2}[F_{2n}(p) - E_{2n}(p)] = \frac{1}{2} \frac{1}{(2n+1)!} [\psi_{2n+1}(p) - \psi_{2n+1}(1-p)];$$

and on substitution from (4.13), (4.14) these become

$$\left. \begin{aligned} B_{2n}(p) &= \frac{1}{2} \frac{1}{(2n)!} \psi_{2n}(p - \frac{1}{2}) \\ B_{2n+1}(p) &= \frac{1}{(2n+1)!} (p - \frac{1}{2}) \psi_{2n}(p - \frac{1}{2}) \end{aligned} \right\} \quad (5.24)$$

The first few functions $B_m(p)$ are

$$\begin{aligned} B_2(p) &= p(p-1)/2 \cdot 2! & B_3(p) &= p(p-\frac{1}{2})(p-1)/3! \\ B_4(p) &= (p+1)p(p-1)(p-2)/2 \cdot 4! \\ B_5(p) &= (p+1)p(p-\frac{1}{2})(p-1)(p-2)/5! \end{aligned}$$

† In this formula, $B_2(p)$ is written for the coefficient of $(\delta^2 f_0 + \delta^2 f_1)$, not for the coefficient of $\mu \delta^2 f_{\frac{1}{2}}$, and similarly for the higher even orders of differences. This usage follows that adopted by Comrie (*Chambers’s 6-Figure Tables*, 1949, vol. 2). In some earlier work and tabulation of coefficients in Bessel’s formulae B'' or B^{II} has been used for the coefficient of $\mu \delta^2 f_{\frac{1}{2}}$.

Bessel's formula can alternatively be derived directly without using Everett's formula; the following is a summary of this derivation.

Expressed in terms of operators, formula (5.23) can be written

$$E^p f_0 = \phi_1(\delta)(E+1)f_0 + \phi_2(\delta)E^{\frac{1}{2}}f_0,$$

where $\phi_1(\delta)$ is an even function of δ and $\phi_2(\delta)$ an odd function. Thus the operators $\phi_1(\delta)$ and $\phi_2(\delta)$ must satisfy

$$E^{p-\frac{1}{2}} = \phi_1(\delta)(E^{\frac{1}{2}} + E^{-\frac{1}{2}}) + \phi_2(\delta),$$

that is

$$e^{(p-\frac{1}{2})U} = \phi_1(\delta) \cdot (2 \cosh \frac{1}{2}U) + \phi_2(\delta).$$

But

$$e^{(p-\frac{1}{2})U} = \cosh(p-\frac{1}{2})U + \sinh(p-\frac{1}{2})U,$$

of which the first term is an even function of U and so of δ , and the second is an odd function. Hence we obtain a formula of the kind sought by taking

$$\phi_1(\delta) = \cosh(p-\frac{1}{2})U / 2 \cosh \frac{1}{2}U, \quad \phi_2(\delta) = \sinh(p-\frac{1}{2})U.$$

The expansions of these in powers of $\delta = 2 \sinh \frac{1}{2}U$ can be written down from (5.16), (5.17) by making the formal substitutions of U for u , δ for z , and $(p-\frac{1}{2})$ for β .

5.42. Use of Bessel's and Everett's formulae

Bessel's formula to second differences, namely,

$$f(x_0 + p \delta x) = f_0 + p \delta f_{\frac{1}{2}} + B_2(p)(\delta^2 f_0 + \delta^2 f_1), \quad (5.25)$$

or to third differences, with second differences modified as explained below, is generally the most useful formula for non-linear interpolation, unless so large a number of figures is required, or the spacing δx is so large, that fourth and perhaps higher-order differences have to be taken into account. Then Everett's formula is probably more convenient, especially when using tables in which only differences of even order are tabulated.

The coefficient $B_2(p)$ is always negative. A critical table to three decimals is given in Comrie and Milne-Thompson's *Standard 4-Figure Tables* and one to four decimals in *Interpolation and Allied Tables* (1956), where $B_3(p)$ and $B_4(p)$ are also tabulated; $B_2(p)$ and $B_3(p)$ are also tabulated in *Chambers's 6-Figure Tables*, vol. 2 (1949). For other tables of coefficients in this (and other) interpolation formulae, reference should be made to the *Index of Mathematical Tables*.

In Bessel's interpolation formula, the coefficients of the odd-order differences are all zero at $p = \frac{1}{2}$ as well as at $p = 0$ and 1; this is an advantage over most other interpolation formulae which involve all orders of differences. The greatest value of $|B_3(p)|$ is about 0.008, so the contribution from $\delta^2 f$ to the interpolated value is less than 0.5 in the least significant figure if $|\delta^3 f|$ is less than 60.

Further, the contribution from the second and fourth differences together is

$$B_2(p)[(\delta^2 f_0 + \delta^2 f_1) + \frac{1}{12}(p+1)(p-2)(\delta^4 f_0 + \delta^4 f_1)]$$

and $-\frac{1}{12}(p+1)(p-2)$ does not vary greatly over the range of p , from 0 to 1, over which this formula will be used; its maximum value is 0.1875 at $p = \frac{1}{2}$ and it is greater than 0.180 over half this range of p ; its smallest value is 0.1667 at $p = 0$ and 1, where $B_2(p)$, by which it is multiplied, is zero. Hence a good approximation to the contribution from $\delta^4 f$ to the interpolated value can be made by subtracting a *constant* multiple of $\delta^4 f$ from each $\delta^2 f$, and applying Bessel's formula, correct to second or third differences only, with the second differences so modified. If we write

$$\delta_m^2 f_j = \delta^2 f_j - C \delta^4 f_j, \quad (5.26)$$

and use $B_2(p)(\delta_m^2 f_0 + \delta_m^2 f_1)$ in such a formula, the residual contribution from $\delta^4 f$ is

$$\frac{1}{2}p(p-1)[C + \frac{1}{12}(p+1)(p-2)]\mu\delta^4 f_{\frac{1}{2}}. \quad (5.27)$$

The best value of C is that which makes the extreme values of the coefficient here equal and opposite, and is $C = 0.184$; the greatest value of the coefficient of $\delta^4 f$ in (5.27) is then 0.00045, whereas the greatest value of $|B_4(p)|$ is 0.0117. The residual contribution from $\delta^4 f$ is less than 0.5 in the least significant figure if $\delta^4 f$ is less than 1100.

Quantities $\delta_m^2 f_j$ given by (5.26) with $C = 0.184$ are called 'modified second differences' and this inclusion of a constant multiple of the fourth differences in modified second differences is called 'throwback' of the fourth differences to the second. It is due to L. J. Comrie, and is a valuable device for simplifying practical interpolation, particularly inverse interpolation and subtabulation.

In Everett's formula E_{2n} , the coefficient of $\delta^{2n} f_0$, is the same function of $(1-p)$ as F_{2n} is of p , so that in tables of interpolation coefficients the number of separate functions which have to be tabulated for Everett's formula is only about half as many as for formulae involving all orders of differences. Also in tables of the function f to be interpolated, only even-order differences need be given. Tables of Everett coefficients are given in *Interpolation and Allied Tables* and in *Chambers's 6-Figure Tables*, vol. 2 (1949); tables at the close interval 0.0001 in p have been published by the Mathematisch Centrum, Amsterdam.†

The 'throwback' can be used with Everett's formula as with Bessel's. The contribution from $\delta^2 f_1$ and $\delta^4 f_1$ together in Everett's formula is

$$\frac{1}{6}p(p^2-1)[\delta^2 f_1 + \frac{1}{20}(p^2-4)\delta^4 f_1];$$

the coefficient $-\frac{1}{20}(p^2-4)$ varies from 0.15 to 0.20 over the range $p = 0$

† *Tables of Everett's Interpolation Coefficients* by E. W. Dijkstra and A. van Wijngaarden (Amsterdam, 1955).

to 1, and is multiplied by a zero factor at both ends of the range. If the same modified second differences are used, namely,

$$\delta_m^2 f = \delta^2 f - 0.184 \delta^4 f,$$

the residual contribution from $\delta^4 f_1$ is

$$\frac{1}{6}p(p^2-1)[0.184 + \frac{1}{20}(p^2-4)]\delta^4 f_1;$$

the greatest value of this coefficient is about 0.0008, so that this contribution is less than 0.3 in the last figure if $\delta^4 f_1$ is less than about 400. Similarly for the contribution from $\delta^4 f_0$.

If fourth differences are too large to be treated by means of the throwback, Everett's formula can be taken as far as the $\delta^4 f$ terms, and the sixth differences thrown back to the fourth differences.† If eighth differences are appreciable, very effective use can be made of a joint throwback of the sixth and eighth differences to the second and fourth differences.†

5.43. Practical details in non-linear interpolation

In using Bessel's or Everett's formulae, values of the coefficients can either be calculated as required or taken from tables. In the latter case the interpolation will have to be done in two stages if the number of decimals in p is greater than that in the argument of the tables of interpolation coefficients. One method of dealing with this situation is to carry out a subsidiary interpolation in the tables of interpolation coefficients themselves. But it is generally better to carry out a small subtabulation of the function $f(x)$ using only tabular values of the interpolation coefficients. For example, if $f(x)$ is tabulated at intervals $\delta x = 0.1$ and its value is wanted for $x = 0.854377$, and available tables of the interpolation coefficients have the argument p at intervals $\delta p = 0.001$, the values of $f(x)$ for $x = 0.8541(0.0001)0.8545$ can be obtained without interpolation in the tables of the interpolation coefficients, and interpolation in this small table of $f(x)$ will then give the result required; linear interpolation will often be adequate at this stage.

In carrying out a non-linear interpolation, it is advisable to carry one guarding figure to avoid accumulation of rounding errors from the various contributions to the interpolated value. For a similar reason, a guarding figure should be kept in the subtabulation mentioned in the previous paragraph. Also it is advisable to retain contributions

† For these and other developments of the idea of the throwback, see *Chambers's 6-Figure Tables*, vol. 2 (1949), p. 533.

greater than 0.2 in the least significant digit from the higher orders of differences.

On this basis:

In Bessel's formula, with throwback of fourth differences to second:

$\delta^3 f$ can be neglected if less than 15

$\delta^4 f$ can be neglected if less than 500.

In Everett's formula, with throwback of fourth differences to second:

$\delta^4 f$ can be neglected if less than 250.

Examples:

(a) Given the following values, to find $f(\frac{2}{3})$:

| x | $f(x)$ | δf | $\delta^2 f$ | $\delta^3 f$ | $(\delta^2 f_0 + \delta^2 f_1)$ |
|------|---------|------------|--------------|--------------|---------------------------------|
| 0.60 | 1.66667 | | | | |
| | | -5377 | | | |
| .62 | .61290 | | 337 | | |
| | | -5040 | | -32 | |
| .64 | .56250 | | 305 | | |
| | | -4735 | | -26 | |
| .66 | .51515 | | 279 | | |
| | | -4456 | | -25 | +533 |
| .68 | .47059 | | 254 | | |
| | | -4202 | | -20 | |
| .70 | .42857 | | 234 | | |
| | | -3968 | | | |
| 0.72 | 1.38889 | | | | |

Here $(\delta x) = 0.02$, $x = \frac{2}{3} = 0.66 + \frac{1}{3}(\delta x)$, $p = \frac{1}{3}$. The contribution from the third difference is just worth taking into account, but the fourth-difference contribution is negligible, even without using the throwback. The value of $B_2(p)$ is $\frac{1}{4}p(p-1) = -\frac{1}{18}$, and that of $B_3(p)$ is $\frac{1}{6}p(p-1)(p-\frac{1}{2}) = +\frac{1}{162}$. Hence we have

$$\begin{aligned}
 f_0 &= 1.51515 \\
 p\delta f_{\frac{1}{3}} &= -0.01485_3 \\
 B_2(p)[\delta^2 f_0 + \delta^2 f_1] &= -0.00029_6 \\
 B_3(p)\delta^3 f_{\frac{1}{3}} &= -0.00000_2 \\
 &\quad \underline{1.49999_9}, \\
 \text{or rounded off, } &1.50000.
 \end{aligned}$$

The guarding figure is written here as a suffix; this is a convenient convention.

Notes: (i) The point of expressing the second-difference contribution in the form $B_2(p) \cdot (\delta^2 f_0 + \delta^2 f_1)$, rather than $\{2B_2(p)\} \mu \delta^2 f_{\frac{1}{3}}$, is clear from this example. If the quantity $\delta^2 f_0 + \delta^2 f_1$ is odd, then in dividing by 2 to obtain $\mu \delta^2 f_{\frac{1}{3}}$ we would either have to round off or keep an extra figure, and this is avoided by incorporating the division by 2 in the factor $B_2(p) = \frac{1}{4}p(p-1)$ by which this quantity is multiplied.

(ii) The function $f(x)$ here is $1/x$, the tabular values being rounded off to five decimals, so that the correct value of $f(\frac{2}{3})$ is 1.5 exactly.

(b) Given the following values, to find $f(\frac{2}{3})$:

| x | $f(x)$ | δf | $\delta^2 f$ | $\delta^3 f$ | $\delta^4 f$ | $-0.184\delta^4 f$ | $\delta_m^2 f$ | $\delta_m^2 f_0 + \delta_m^2 f_1$ |
|------|---------|------------|--------------|--------------|--------------|--------------------|----------------|-----------------------------------|
| 0.50 | 2.00000 | | | | | | | |
| | | -18182 | | | | | | |
| .55 | 1.81818 | | 3031 | | | | | |
| | | -15151 | | -701 | | | | |
| .60 | .66667 | | 2330 | | 203 | -37 | 2293 | |
| | | -12821 | | -498 | | | | |
| .65 | .53846 | | 1832 | | 131 | -24 | 1808 | |
| | | -10989 | | -367 | | | | 3256 |
| .70 | .42857 | | 1465 | | 93 | -17 | 1448 | |
| | | -9524 | | -274 | | | | |
| .75 | .33333 | | 1191 | | 63 | -12 | 1179 | |
| | | -8333 | | -211 | | | | |
| .80 | .25000 | | 980 | | | | | |
| | | -7353 | | | | | | |
| 0.85 | 1.17647 | | | | | | | |

Here $(\delta x) = 0.05$, $x = \frac{2}{3} = 0.65 + \frac{1}{3}(\delta x)$, so $p = \frac{1}{3}$, $q = \frac{2}{3}$.

(i) By Bessel's formula

$$\begin{aligned} B_2(p) &= \frac{1}{4}p(p-1) = -\frac{1}{18} \\ B_3(p) &= \frac{1}{6}p(p-1)(p-\frac{1}{2}) = +\frac{1}{162} \\ f_0 &= 1.53846 \\ p\delta f_{\frac{1}{2}} &= -0.03663_0 \\ B_2(p)(\delta_m^2 f_0 + \delta_m^2 f_1) &= -0.00180_9 \\ B_3(p)\delta^3 f_{\frac{1}{2}} &= -0.00002_3 \\ &\quad \underline{1.49999_8} \end{aligned}$$

(ii) By Everett's formula

$$\begin{aligned} E_2(p) &= -\frac{1}{6}q(1-q^2) = -\frac{5}{81} \\ F_2(p) &= -\frac{1}{6}p(1-p^2) = -\frac{4}{81} \\ f_0 &= 1.53846 \\ p\delta f_{\frac{1}{2}} &= -0.03663_0 \\ E_2(p)\delta_m^2 f_0 &= -0.00111_6 \\ F_2(p)\delta_m^2 f_1 &= -0.00071_5 \\ &\quad \underline{1.49999_9} \end{aligned}$$

Rounded off to five decimals = 1.50000 Rounded off to five decimals
= 1.50000

In this example modified second differences have been used, and the residual contributions from fourth differences are negligible. If modified second differences had not been used, it would have been necessary to include the fourth difference terms in each case.

(iii) By preliminary subtabulation:

| p | .30 | .32 | .34 | .36 |
|-------------------------------------|----------------------------|----------------------------|----------------------------|----------------------------|
| $E_2(p) = \frac{1}{6}q(1-q^2)$ | .05950 | .06093 | .06208 | .06298 |
| $F_2(p) = \frac{1}{6}p(1-p^2)$ | .04550 | .04787 | .05012 | .05222 |
| x | .665 | .666 | .667 | .668 |
| f_0 | 1.53846 | 1.53846 | 1.53846 | 1.53846 |
| $p\delta f_{\frac{1}{2}}$ | -3296 ₇ | -3516 ₅ | -3736 ₃ | -3956 ₀ |
| $\frac{1}{6}q(1-q^2)\delta_m^2 f_0$ | -107 ₆ | -110 ₂ | -112 ₂ | -113 ₉ |
| $\frac{1}{6}p(1-p^2)\delta_m^2 f_1$ | -65 ₉ | -69 ₃ | -72 ₆ | -75 ₆ |
| f | <u>1.50375₈</u> | <u>1.50150₀</u> | <u>1.49924₉</u> | <u>1.49700₅</u> |

Differences at the smaller interval $\left\{ \begin{array}{cccc} -225_8 & & -225_1 & \\ & 0_7 & & 0_7 \\ & & -224_4 & \end{array} \right.$

Linear interpolation between the subtabulated values is now adequate; $p = \frac{2}{3}$, so

$$\begin{aligned} f(\frac{2}{3}) &= 1.50150_0 - \frac{2}{3}(225_1) \\ &= 1.50000 \text{ on rounding off to five decimals.} \end{aligned}$$

Note: Since the interval length has been reduced by a factor 50, second differences are reduced by a factor 2500 from those of the original table, so their values would be expected to be about 0_8 in the fifth decimal, and are certainly negligible for interpolation purposes. It is then only necessary to calculate two values of f (for example those for $x = 0.666$ and 0.667); but four values have been calculated to give a partial check. A thorough check of an isolated interpolation is difficult to achieve, but a good check is provided by carrying out the interpolation between two sets of values of the original function at different intervals.

5.5. Lagrange's formula

The interpolation formulae so far given have expressed the interpolated value of $f(x)$ in terms of contributions from its various orders of differences. An alternative type of formula expresses the interpolated value of $f(x)$ as a sum of multiples of the values of the function f itself, with coefficients which are functions of the fraction p of the interval length for which the interpolation is required, thus:

$$f(x_0 + p \delta x) = \sum_j L_j(p) f_j. \quad (5.28)$$

An interpolation formula of this type is called a 'Lagrange interpolation formula', and the coefficients $L_j(p)$ are known as 'Lagrange interpolation coefficients'. There are several formulae of this type, with different numbers of terms taken in the sum in (5.28), and correspondingly with different sets of coefficients.

A formula using n function values is usually known as an ' n -point' formula; n is usually taken as even, and an equal number of points taken on each side of the interval in which interpolation is to be carried out. Such formulae can be obtained by expressing the finite differences in Bessel's or Everett's formula in terms of function values by formula (4.8) and collecting terms involving the same function value. But they are more conveniently obtained as special cases of a formula, which will be derived in § 5.7, for interpolation of a function given at unequal intervals of the argument. An n -point formula is based on the approximation to f by a polynomial of degree $(n-1)$ through n successive values of the function, interpolation being carried out by evaluating this polynomial at the value of x for which the interpolation is required. For even values of n , use of such a formula is equivalent to the use of Bessel's formula to $(n-1)$ th differences or of Everett's to $(n-2)$ th differences, without throwback.

Tables of Lagrange coefficients for 4-point and 6-point interpolation, for $p = 0(0.01)1.00$, are given in *Chambers's 6-Figure Tables*; for other tabulations the *Index of Mathematical Tables* should be consulted.

The advantage of Lagrangian coefficients formulae, if it is an advantage, is that they can be used directly on tables in which no differences are given. On the other hand, they have several disadvantages, as pointed out by Comrie:†

- (i) They provide no check that the function values used in them have been taken correctly, whereas the differences used in a difference formula also provide a check on the function values used;
- (ii) A single calculation of an interpolated value provides no indication whether the degree of the polynomial used is inadequate, adequate, or excessive;
- (iii) At least one of the coefficients $L_j(p)$ is greater than 0.5; if therefore an interpolation is required for a value of p which is not a tabular value in the table of Lagrangian interpolation coefficients, interpolation in these tables is required to the same number of significant figures as that required for the interpolation of f itself;
- (iv) Use of Lagrangian formulae does not lend itself to an easy process for inverse interpolation.

To these may be added:

- (v) They do not provide the facilities for improving the accuracy of interpolation without complicating the formulae, such as are provided by interpolation formulae in terms of differences by use of the 'throwback'. The Lagrangian formulae are based on the approximation to the function to be interpolated by a polynomial of the n th degree through $(n+1)$ points. But consider the significance of the use of the throwback from fourth to second differences and subsequent use of Everett's formulae to second differences. The fact that this modification of the second differences improves the accuracy of the interpolation (and moreover by a factor of about 10, not only by a small amount) means that for interpolation between f_0 and f_1 the best cubic is *not* the cubic through f_{-1}, f_0, f_1, f_2 which is the one used in the four-point Lagrange formula.

Certainly there is a formula of Lagrangian type corresponding to the Everett formula to second differences, used with modified second differences; but this is no simpler than a six-point Lagrangian formula based on the use of a quintic polynomial. The corresponding formula using differences is Everett's to fourth differences, and if in this the joint throwback of sixth and eighth

† *Chambers's 6-Figure Tables*, vol. 2 (1949), Introduction, p. xxix.

differences to second and fourth is used, a very substantial improvement in the accuracy, compared with that of a six-point Lagrangian interpolation, is achieved.

Comrie's comment† is that he 'has to admit that his experience has not made him partial to blind Lagrangian interpolation, except when special circumstances point very definitely to it'.

5.51. Special interpolation methods for particular functions

For some particular functions, special interpolation methods may be more convenient than the use of the Bessel or Everett formulae. For example, for the exponential function it may be most convenient to use the addition formula

$$e^{x+y} = e^x e^y$$

and carry out interpolation by a multiplication or succession of multiplications. If, for example, e^x is tabulated at intervals of 0.01 in x , an auxiliary table for $x = 0(0.0001)0.01$ would enable values to be obtained for four-decimal values of x in the range of the main table by a single multiplication; alternatively a set of auxiliary tables for $x = [0(0.1)1] \times 10^{-n}$ for n from 2 to 5, say, would enable values of e^x for six-decimal values of x to be obtained with not more than four multiplications.

If the function y to be interpolated satisfies a simple differential equation such as $y'' = xy$, formulae for successive derivatives of y , obtained by successive differentiation of the differential equation, may be simple enough to be used for the numerical evaluation of these derivatives. Then Taylor's series can be used for interpolation between tabular values, $f(x_0 + p\delta x)$ being calculated from as many terms of the series

$$f(x_0 + p\delta x) = f_0 + p(\delta x)f'_0 + \frac{1}{2!}p^2(\delta x)^2f''_0 + \frac{1}{3!}p^3(\delta x)^3f'''_0 + \dots \quad (5.29)$$

as are appreciable. A good check is provided by evaluation of the alternative expansion, in terms of f and its derivatives at $x = x_1$, namely:

$$\begin{aligned} f(x_0 + p\delta x) &= f(x_1 - q\delta x) \quad (\text{where } q = 1 - p) \\ &= f_1 - q(\delta x)f'_1 + \frac{1}{2!}q^2(\delta x)^2f''_1 - \frac{1}{3!}q^3(\delta x)^3f'''_1 + \dots \quad (5.30) \end{aligned}$$

The convenient quantities to tabulate for interpolation purposes are not the derivatives $f^{(k)}$ but the quantities $[(\delta x)^k/k!]f^{(k)}$, sometimes called 'reduced derivatives'. This method of interpolation is particularly convenient in the case of functions which have been evaluated by integration of the appropriate differential equation by the Taylor series

† *Chambers's 6-Figure Tables*, vol. 2 (1949), Introduction, p. xxx.

method (§ 7.4), since the reduced derivatives are evaluated in the course of this calculation, and so are available for interpolation purposes without any further work. An example is provided by the tables† of the function usually written $\text{Bi}(x)$, which is one solution of the equation $y'' = xy$.

5.6. Subtabulation

Subtabulation is a special case of interpolation, of which the purpose is to take a function $f(x)$ at tabular interval δx and construct from it a table at a smaller tabular interval $s\delta x$; in practice s is usually $\frac{1}{2}$, $\frac{1}{5}$, or $\frac{1}{10}$. The values of the function at the large interval (δx), between which interpolation is carried out, are called 'pivotal values' of $f(x)$.

It is a valuable process when the direct calculation of $f(x)$ is difficult or long, for example by summing a series of a large number of terms, or evaluation of a definite integral in which the integrand is a function of x . In such cases we want to restrict the number of values of x for which $f(x)$ is calculated directly, and derive from them a table at a smaller interval, probably one in which second differences at most have to be included in interpolation, and possibly one in which linear interpolation is adequate. There is no purpose in subtabulation if linear interpolation is already adequate; the purpose of subtabulation is to break down the tabular interval when linear interpolation is certainly not adequate; so it is essentially concerned with non-linear interpolation.

Since $\delta^n f$ is of order $(\delta x)^n$, the higher-order differences are very much reduced by even a moderate degree of subtabulation. For example, subtabulation to fifths reduces fourth differences by a factor of over 500 and sixth differences by a factor of over 15,000.

In subtabulation, a systematic set of results is required, instead of an isolated result as is more usual in an interpolation process. This suggests that a systematic procedure should be used for obtaining the results.

We could interpolate a sequence of values of the function itself and check the differences; alternatively we could construct the sequence of second (or higher) order differences of the function at the smaller interval and build up from these, using the facilities for building up a function from its differences (the 'National' machine, for example, if available). The latter process has the advantages that (i) most of the work is done with small numbers, and (ii) a good overall check is provided by the reproduction of the pivotal values in the course of the summation of the

† *British Association Mathematical Tables*, Part-volume B, *The Airy Integral* (1946).

differences of the function at the smaller interval. The former does not check that the correct pivotal values have been taken, and the use of incorrect pivotal values may not be apparent in the differences of the subtabulated function; these differences check against random errors, but the effect of an incorrect pivotal value is a systematic error, and may not be indicated by differences. This is illustrated by the following example.

Consider the subtabulation to tenths of $\sin x$ from a table at intervals of 10° . Use of an incorrect value at $x = 50^\circ$ (0.76624 for 0.76604) might give a set of subtabulated values as follows:

| x | ' $\sin x$ ' | | | x | ' $\sin x$ ' | | |
|-----|----------------|------|-----|-----|----------------|------|-----|
| 30° | <u>0.50000</u> | 1504 | | 50° | <u>0.76624</u> | 1110 | -23 |
| 31 | 1504 | 1488 | -16 | 51 | 7734 | 1086 | -24 |
| 32 | 2992 | 1472 | -16 | 52 | 8820 | 1061 | -25 |
| 33 | 4464 | 1455 | -17 | 53 | 0.79881 | 1036 | -25 |
| 34 | 5919 | 1438 | -17 | 54 | 0.80917 | 1010 | -26 |
| 35 | 7357 | 1420 | -18 | 55 | 1927 | 986 | -24 |
| 36 | 0.58777 | 1402 | -18 | 56 | 2913 | 961 | -25 |
| 37 | 0.60179 | 1385 | -17 | 57 | 3874 | 935 | -26 |
| 38 | 1564 | 1367 | -18 | 58 | 4809 | 910 | -25 |
| 39 | 2931 | 1348 | -19 | 59 | 5719 | 884 | -26 |
| 40 | 0.64279 | 1329 | -19 | 60 | 0.86603 | 858 | -26 |
| 41 | 5608 | 1309 | -20 | 61 | 7461 | 832 | -26 |
| 42 | 6917 | 1289 | -20 | 62 | 8293 | 806 | -26 |
| 43 | 8206 | 1269 | -20 | 63 | 9099 | 780 | -26 |
| 44 | 0.69475 | 1248 | -21 | 64 | 0.89879 | 752 | -28 |
| 45 | 0.70723 | 1226 | -22 | 65 | 0.90631 | 724 | -28 |
| 46 | 1949 | 1203 | -23 | 66 | 1355 | 695 | -29 |
| 47 | 3152 | 1181 | -22 | 67 | 2050 | 668 | -27 |
| 48 | 4333 | 1158 | -23 | 68 | 2718 | 640 | -28 |
| 49 | 5491 | 1133 | -25 | 69 | 3358 | 611 | -29 |
| 50 | <u>0.76624</u> | | -23 | 70 | <u>0.93969</u> | | |

Here the pivotal values are underlined; that at $x = 50^\circ$ is in error by 20 units in the fifth decimal, the others are correct to five decimals. The differences of the subtabulated values are no more irregular than would be expected as the result of rounding errors, and certainly contain no suggestion of an error of 20 units.

Certainly such an error in a pivotal value ought to be detected by differencing the pivotal values before beginning the subtabulation. But if a Lagrangian formula were used for carrying out the subtabulation, this step might be omitted since the differences of the pivotal values would not be used in the subtabulation process; if they were to be obtained to check the pivotal values, it would be better to use them in

the subtabulation also. The point of this example, however, is to show that differencing the subtabulated values does not *by itself* provide an adequate check of the subtabulation process.

5.61. End-figure method of subtabulation

Comrie† has given a convenient process for subtabulation, in which only the *last digit* of each interpolated value is evaluated by the use of a suitable interpolation formula, and the complete values are then built up from their differences. From the last digits in the interpolated function values, only the last digits in the differences can be obtained directly; but in subtabulation at a fraction s of the interval between the pivotal values, the n th differences of the subtabulated values are approximately s^n times those of the pivotal values, and for some value of n these n th differences of the subtabulated values will vary slowly enough for their last digits to establish the values of the differences themselves. Suppose for example that, for the pivotal values, $\delta^2 f_j = 610$ and $\delta^2 f_{j+1} = 505$, and subtabulation is to fifths ($s = \frac{1}{5}$). Then the second differences of the subtabulated values will be approximately $\frac{1}{25}$ of those of the pivotal values, that is (allowing for possible effects of rounding errors) from about 25 at the beginning of this interval to 19 at the end. Hence if the last digits of the second differences of the interpolated values are

$$5, 3, 3, 1, 0,$$

these second differences can with confidence be given the values

$$25, 23, 23, 21, 20,$$

and the function can then be evaluated by summation from these values (see § 4.46). If any mistake is made, it is shown up by the pivotal values not being reproduced in the summation.

The process can be illustrated in a simple case by example (b) of § 5.43 (see p. 73). If in that example the values of $f(0.665)$ and $f(0.666)$ had been evaluated in full, but only the last digits of $f(0.667)$ and $f(0.668)$ had been determined, the results (rounded off to five decimals) would have been:

| | | | | |
|--------|---------|---------|--------|--------|
| x | 0.665 | 0.666 | 0.667 | 0.668 |
| $f(x)$ | 1.50376 | 1.50150 |5 |0 |

| | | | | |
|--|---|-------|-------|-------|
| Differences at the smaller interval | } | — 226 | — ..5 | — ..5 |
| | | | | |

† L. J. Comrie, *Monthly Notices, R.A.S.*, 88 (1928), 506; *Interpolation and Allied Tables*, incorporated in *Nautical Almanac*, 1931.

where the dots represent digits so far undetermined. But, as mentioned on p. 74, the second differences of f at the smaller interval are about 0_8 so each of the first differences whose last digit is 5 must be -225 . The values of $f(x)$ at $x = 0.667$ and 0.668 could then be built up from these differences.

The following is a more extensive example, in which it is necessary to go to second differences before writing down the values of a set of differences from their last digits, and which also illustrates some further points of procedure.

Example: Given the following table:

| x | $f(x)$ | | | |
|-------|---------|-------|-----|----|
| 0.310 | 2.96671 | | | |
| | | 8476 | | |
| 0.305 | 3.05147 | | 340 | |
| | | 8816 | | 20 |
| 0.300 | 3.13963 | | 360 | |
| | | 9176 | | 21 |
| 0.295 | 3.23139 | | 381 | |
| | | 9557 | | 22 |
| 0.290 | 3.32696 | | 403 | |
| | | 9960 | | 26 |
| 0.285 | 3.42656 | | 429 | |
| | | 10389 | | |
| 0.280 | 3.53045 | | | |

to subtabulate to fifths (i.e. at intervals 0.001 in x) from $x = 0.300$ to $x = 0.290$.

(*Note:* for this function it is convenient to take differences in the direction of x decreasing, as in this table; then, apart from effects of rounding errors, they are all positive.)

Everett's formula to second differences (unmodified) is adequate in this case; the Everett coefficients for the points of subtabulation are:

| p | 0 | $\frac{1}{5}$ | $\frac{2}{5}$ | $\frac{3}{5}$ | $\frac{4}{5}$ | 1 |
|----------|---|---------------|---------------|---------------|---------------|---|
| $E_2(p)$ | 0 | $-\cdot 048$ | $-\cdot 064$ | $-\cdot 056$ | $-\cdot 032$ | 0 |
| $F_2(p)$ | 0 | $-\cdot 032$ | $-\cdot 056$ | $-\cdot 064$ | $-\cdot 048$ | 0 |

The second differences of the pivotal values over the range specified for the subtabulation are from 360 to 403, so for subtabulation to fifths the second differences of the subtabulated values will be from about 14 to 16. Thus it is only necessary to difference the end digits to second differences in order to be sure of the values to be used in building up the function values. Since this building up is to be done from second differences, we will need two function values from which to start, namely the pivotal value $f(0.300)$ and the value $f(0.299)$. For the latter, Everett's formula gives:

$$\begin{aligned}
 & f(0.300) && 3.13963 \\
 & + \frac{1}{5} \delta f(0.2995) && 1835_2 \\
 & + E_2(\frac{1}{5}) \delta^2 f(0.300) && -17_3 \\
 & + F_2(\frac{1}{5}) \delta^2 f(0.295) && -12_2 \\
 & f(0.299) = && \underline{3.15768_7}, \text{ or, rounded off, } 3.15769.
 \end{aligned}$$

Only the rounded value is needed subsequently; guarding figures are kept in the contributions to it, but can be discarded in the interpolated value itself. Guarding figures, written as suffixes, will similarly be kept in the contributions to the other subtabulated values.

For the interval $x = 0.300$ to 0.295 , the last two digits of δf for the pivotal values are 76, so the first differences of the linear contributions $f_0 + p\delta f_1$ to the interpolated values in this interval end with 5₂. The end figure of $f(0.300)$ is 3, and successive additions of 5₂ give for the end figures of $f(0.300)$ to $f(0.295)$ inclusive the values

$$3_0, 8_2, 3_4, 8_6, 3_8, 9_0$$

of which the last is the end figure of the pivotal value $f(0.295)$; this provides a check on the additions. For the next interval the first difference of the pivotal values ends with 57, so the first differences of the linear contributions to the interpolated values end with 1₄, and for $f(0.295)$ to $f(0.290)$ inclusive are

$$9_0, 0_4, 1_8, 3_2, 4_6, 6_0;$$

the comparison of the last of these with the pivotal value $f(0.290)$ again furnishes a check.

The complete calculation can be arranged in tabular form as follows (δ is used for the central-difference operator at the smaller intervals):

| x | $\delta^2 f_0$ | $\delta^2 f_1$ | $\frac{1}{2}\delta f$ | (a) | (b) | (c) | (d) | Last figure of f | | $\delta^2 f$ | δf | f |
|-------|----------------|----------------|-----------------------|----------------|------------------|------------------|------------------|--------------------------|---|--------------|------------|---------|
| 0.300 | 360 | 381 | 5 ₂ | 3 ₀ | | | = 3 ₀ | 3 | | | | 3.13963 |
| 299 | | | | 8 ₂ | -17 ₃ | -12 ₂ | = 8 ₇ | 9 | 6 | | 1806 | 5769 |
| 298 | | | | 3 ₄ | -23 ₀ | -21 ₃ | = 9 ₁ | 9 | 0 | 4 | 1820 | 7589 |
| 297 | | | | 8 ₆ | -20 ₂ | -24 ₄ | = 4 ₀ | 9 | 5 | 5 | 1835 | 19424 |
| 296 | | | | 3 ₈ | -11 ₅ | -18 ₃ | = 4 ₀ | 4 | 0 | 5 | 1850 | 3.21274 |
| 295 | | | | 9 ₀ | | | = 9 ₀ | 4 | 5 | 5 | 1865 | 3139 |
| 294 | 381 | 403 | 1 ₄ | 0 ₄ | -18 ₃ | -13 ₀ | = 9 ₁ | 9 | 0 | 5 | 1880 | 5019 |
| 293 | | | | 1 ₈ | -24 ₄ | -22 ₇ | = 4 ₇ | 9 | 6 | 6 | 1896 | 6915 |
| 292 | | | | 3 ₂ | -21 ₃ | -26 ₀ | = 5 ₉ | 5 | 1 | 5 | 1911 | 28826 |
| 291 | | | | 4 ₆ | -12 ₂ | -19 ₅ | = 2 ₉ | 6 | 6 | 6 | 1927 | 3.30753 |
| | | | | | | | | 3 | 7 | 6 | 1943 | |
| 0.290 | | | | 6 ₀ | | | = 6 ₀ | 3 | 3 | | | 3.32696 |

(a) Linear contribution $f_0 + p\delta f_1$, end figure and guarding figure only.

(b), (c): $E_2(p)\delta^2 f_0$ and $F_2(p)\delta^2 f_1$.

(d) Sum of (a), (b), (c), end figure and guarding figure only.

The column headed $\frac{1}{2}\delta f$ gives the last figure and guarding figure, as required for building up the linear contributions to the interpolated values in column (a). Columns (b) and (c) give the second-difference contributions to the interpolated values; these are here given in full as this makes the work easier to follow; however only the last full digit and a guarding figure are necessary. The sums of entries

(a), (b), (c) with guarding figures are given in column (d) to illustrate the procedure, but only the rounded values of the last digit, as given in the next column, are required for the subsequent work. These rounded values are then differenced to second differences, and from the result we have already had, that the values of $\delta^2 f$ are about 14 to 16, the complete values can now be written down from their last digits. Then from $f(0.300)$, $f(0.299)$ and these second differences, the values of f can be built up. A thorough check is provided by the reproduction of all the pivotal values.

Notes: (i) In this example, $f(x)$ is the function $[(1/x)^\gamma - 1]/\gamma$ with $\gamma = 1.4$. The pivotal values were calculated from this defining formula; for the intermediate values subtabulation is a much quicker and easier process than evaluation of this formula.

(ii) It is not necessary to carry out the full evaluation of $f(0.299)$ by interpolation. The value of $\delta f(0.2995)$ is approximately $\frac{1}{10}[f(0.305) - f(0.295)]$ which is 1799—rather larger since $\delta^2 f$ is positive—and its last digit is 6, which indicates the value 1806 with some certainty. But even if a wrong value were taken, this would be shown by the pivotal value $f(0.295)$ not being reproduced; and from the amount of the discrepancy, the corrections to be made can easily be determined. For example, if $\delta f(0.2995)$ were taken as 1796 instead of 1806, each of the first five first differences would be in error by -10 , so the discrepancy between the value of $f(0.295)$ obtained by building up and the pivotal values would be -50 ; this would indicate that each of these first differences must be increased by 10. (This correction process is not available if the subtabulated function values are built up from differences of higher order than the second.)

(iii) If the second differences of the subtabulated values vary too rapidly for their last figures to be a certain indication of their complete values, a corresponding process involving building up from third or fourth differences can be used. Alternatively, the process could be carried out with the last *two* digits in each subtabulated value instead of with the last digit only.

Another method of subtabulation, also suggested by Comrie,[†] involves the direct calculation of second or fourth differences of the subtabulated values from formulae relating differences at interval $s\delta x$ to differences at interval δx . The subtabulated values are then built up from their differences.

5.7. Interpolation of a function given at unequal intervals of the argument

For the interpolation of a function given at unequal intervals of the argument, the interpolation formula usually used is that of Lagrange. This is based on the use of an n th degree polynomial which takes the given function values at $(n+1)$ values of x . Such a formula is called an $(n+1)$ -point formula. An even number of points (odd value of n) gives an equal number on either side of the value of x for which the interpolation is to be carried out.

[†] L. J. Comrie, *Journ. Roy. Stat. Soc.*, Supplement 3 (1936), 87.

If the function values are f_0, f_1, \dots, f_n at $x = x_0, x_1, \dots, x_n$, not necessarily equally spaced, the polynomial of lowest degree which takes these values is

$$F = \left(\frac{x-x_1}{x_0-x_1} \frac{x-x_2}{x_0-x_2} \dots \frac{x-x_n}{x_0-x_n} \right) f_0 + \left(\frac{x-x_0}{x_1-x_0} \frac{x-x_2}{x_1-x_2} \dots \frac{x-x_n}{x_1-x_n} \right) f_1 + \dots; \quad (5.31)$$

this polynomial is of course not the function f itself, unless f is a polynomial of degree n or lower; it is the polynomial which coincides with f at $x = x_0, x_1, \dots, x_n$. Interpolation is done by evaluating this polynomial for the intermediate value of x .

If f is a polynomial of degree n or less, then $F = f$ and the interpolation formula (5.31) is closely related to the expansion in partial fractions of $f/(x-x_0)(x-x_1)\dots(x-x_n)$. For if $g(x) = 0$ has roots $x = x_0, x_1, \dots, x_n$, all distinct, the expansion of $f(x)/g(x)$ in partial fractions is

$$\frac{f(x)}{g(x)} = \sum_j \frac{f(x_j)}{g'(x_j)(x-x_j)}$$

and this, applied to $g(x) = (x-x_0)(x-x_1)\dots(x-x_n)$, is

$$\frac{f}{(x-x_0)(x-x_1)\dots(x-x_n)} = \frac{f_0}{(x_0-x_1)(x_0-x_2)\dots(x_0-x_n)} \frac{1}{x-x_0} + \dots$$

which is just another form of formula (5.31).

Lagrange's interpolation formula (5.31) is not restricted to functions given at unequal intervals of the argument. Its application when the intervals of x are equal has already been mentioned in §§ 5.5 and 5.6, and its disadvantages in that context pointed out. For functions not tabulated at equal intervals, however, some form of it may be the only method available. It is then important to systematize the work of evaluating the polynomial, since if it is not done in a systematic way it is easy to make a mistake, and adequate checking is at best difficult. One scheme of working, in which the coefficients of the various values of f_j in formula (5.31) are first calculated and checked and are then used to form the sum (5.31), has been given by Comrie;† an important feature is the check of the coefficients which is provided. In another type of method, suggested by Aitken,‡ the result is obtained by a sequence of steps each of which is similar to a linear interpolation; this is considered in § 5.71. Another way of arranging the work is considered in § 5.72.

† L. J. Comrie, *Chambers's 6-Figure Tables*, vol. 2 (1949), Introduction, p. xxxi.

‡ A. C. Aitken, *Proc. Edin. Math. Soc.*, ser. 2, 3 (1932), 56.

5.71. Evaluation of Lagrange's interpolation formula by a sequence of linear cross-means

Linear interpolation or extrapolation of $f(x)$ from the values f_a, f_b , of f at $x = x_a, x_b$ respectively, gives

$$f(x) = \frac{(x_b - x)f_a + (x - x_a)f_b}{x_b - x_a}. \quad (5.32)$$

Aitken calls this quantity the 'linear cross-mean' between f_a and f_b ; let it be written $f_{a,b}(x)$. Linear interpolation between the values $f_{a,b}(x)$

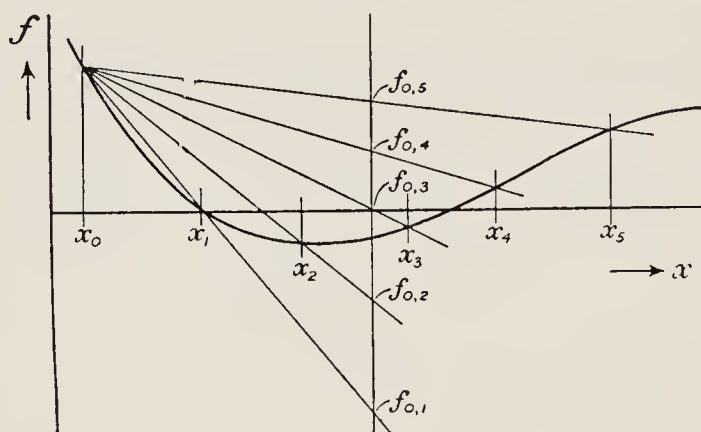


FIG. 6.

and $f_{b,c}(x)$, regarded as values of an auxiliary function at $x = x_a$ and $x = x_c$ respectively, gives

$$f(x) = \frac{(x_c - x)f_{a,b}(x) + (x - x_a)f_{b,c}(x)}{x_c - x_a}.$$

Let this be written $f_{a,b,c}(x)$. It can be verified that this is the value of $f(x)$ given by a three-term Lagrange interpolation formula using the values of f at $x = x_a, x_b$, and x_c .

In general, let $f_{a,b,c,\dots,j,k,l}(x)$ be a set of numbers obtained by successive use of the linear cross-mean formula

$$f_{a,b,c,\dots,i,j,k}(x) = \frac{(x_k - x)f_{a,b,c,\dots,i,j}(x) + (x - x_a)f_{b,c,\dots,i,j,k}(x)}{x_k - x_a}. \quad (5.33)$$

(n suffixes)

Then it can be proved by induction that $f_{a,b,c,\dots,i,j,k}(x)$ is the value of $f(x)$ given by an n -point Lagrangian interpolation formula using the values of f at $x = x_a, x_b, \dots, x_j, x_k$, which need not be in monotonic sequence.

The process suggested by Aitken consists of forming the linear cross-means $f_{0,1}(x), f_{0,2}(x), f_{0,3}(x), \dots$ and in general $f_{0,j}(x)$, then using these to form $f_{0,1,2}(x), f_{0,1,3}(x), \dots$ and in general $f_{0,1,j}(x)$, then $f_{0,1,2,3}(x), f_{0,1,2,4}(x), \dots, f_{0,1,2,j}(x)$ and so on. A graphical representation of the formation of the first

set of linear cross-means $f_{0,j}(x)$ is shown in Fig. 6. An alternative order of procedure† consists of forming first $f_{0,1}(x)$, $f_{1,2}(x)$, ... and in general $f_{j,j+1}(x)$, then $f_{0,1,2}(x)$, $f_{1,2,3}(x)$, ..., $f_{j,j+1,j+2}(x)$ and so on. The formation of the first set of linear cross-means $f_{j,j+1}(x)$ in this procedure is shown in Fig. 7.

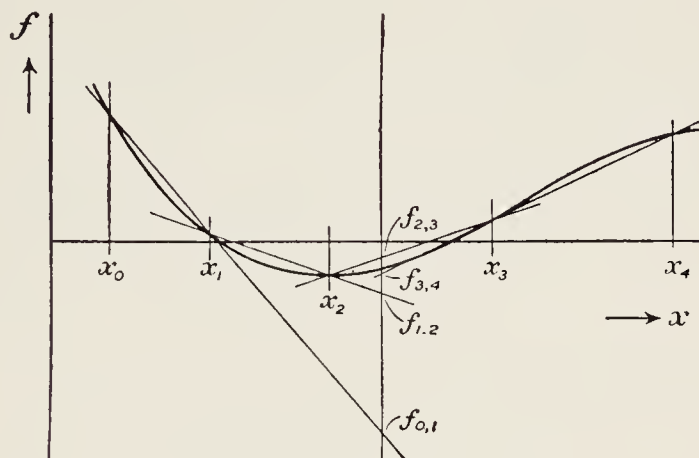


FIG. 7.

In using the latter procedure, the work can be arranged as follows:

| x | $x - x_j$ | f_j | $f_{j,j+1}(x)$ | $f_{j,j+1,j+2}(x)$ | $f_{j,j+1,j+2,j+3}(x)$ |
|-------|-----------|-------|----------------|--------------------|------------------------|
| x_0 | $x - x_0$ | f_0 | | | |
| x_1 | $x - x_1$ | f_1 | $f_{0,1}(x)$ | $f_{0,1,2}(x)$ | |
| x_2 | $x - x_2$ | f_2 | $f_{1,2}(x)$ | $f_{1,2,3}(x)$ | $f_{0,1,2,3}(x)$ |
| x_3 | $x - x_3$ | f_3 | $f_{2,3}(x)$ | $f_{2,3,4}(x)$ | $f_{1,2,3,4}(x)$ |
| x_4 | $x - x_4$ | f_4 | $f_{3,4}(x)$ | | |

This layout is similar to that of a difference table, though the entries are not differences. The entries used in forming $f_{1,2,3,4}(x)$ are enclosed in 'boxes', and the way in which they are selected is shown by the lines joining them to the entry $f_{1,2,3,4}(x)$.

This process has the advantages

- (i) successive calculations are all repetitions of this simple process of forming linear cross-means;
- (ii) the results provide their own criterion of when the process has been carried far enough;
- (iii) common leading figures in two of the linear cross-means can be suppressed in taking higher cross-means.

† E. H. Neville, *Journ. Indian Math. Soc.* **20** (1934), 87. This paper includes an extension of the procedure to use known values of derivatives of $f(x)$.

linear combination of a set of function values with coefficients which are functions of the x_j 's only. It can easily be proved by induction that if $j \leq k \leq j+n$, the coefficient of f_k in $f(x_j, x_{j+1}, \dots, x_{j+n})$ is

$$1 / \prod_{i=j}^{j+n}{}' (x_k - x_i),$$

the dash in Π' indicating that the factor with $i = k$ is omitted from the product; if k is outside the range j to $j+n$, the coefficient of f_k is zero. The value of this coefficient is unaltered by a change in the order of the factors, and it follows that for any function $f(x)$ the value of a divided difference $f(x_j, x_{j+1}, \dots, x_{j+n})$ depends only on the values of x_j involved, and not on the order in which they are taken.

Now consider the function $f(x) = x^n$. The first-order divided differences are given by

$$\frac{f(x_{j+1}) - f(x_j)}{x_{j+1} - x_j} = x_{j+1}^{n-1} + x_j x_{j+1}^{n-2} + \dots + x_j^{n-2} x_{j+1} + x_j^{n-1},$$

a homogeneous polynomial of degree $(n-1)$ in x_j, x_{j+1} . Similarly $f(x_j, x_{j+1}, x_{j+2})$ is a homogeneous polynomial of degree $(n-2)$; and by induction it can be shown that $f(x_j, x_{j+1}, \dots, x_{j+m})$ is a polynomial of degree $n-m$. For consider the difference

$$f(x_{j+1}, \dots, x_{j+k}, x_{j+k+1}) - f(x_j, x_{j+1}, \dots, x_{j+k}). \quad (5.36)$$

Since divided differences are independent of the order in which the values of x_j are taken, this difference is

$$f(x_{j+k+1}, x_{j+1}, \dots, x_{j+k}) - f(x_j, x_{j+1}, \dots, x_{j+k}),$$

and this is zero if $x_{j+k+1} = x_j$. Hence if the k th order divided difference $f(x_j, x_{j+1}, \dots, x_{j+k})$ is a polynomial in x_j, \dots, x_{j+k} , the difference (5.36) contains $(x_{j+k+1} - x_j)$ as a factor, and the divided difference $f(x_j, x_{j+1}, \dots, x_{j+k}, x_{j+k+1})$ is a polynomial of degree one lower than $f(x_j, x_{j+1}, \dots, x_{j+k})$. Now for $f(x) = x^n$, $f(x_j, x_{j+1})$ is a polynomial of degree $(n-1)$, so $f(x_j, x_{j+1}, x_{j+2})$ is a polynomial of degree $(n-2)$, and so on. In particular $f(x_j, x_{j+1}, \dots, x_{j+n})$ is a polynomial of degree zero, that is a constant; it is therefore independent of the values of x_j, \dots, x_{j+n} , and has the same value as if these were equally spaced, namely, 1.

For a polynomial of the n th degree with leading term $a_0 x^n$, the n th-order divided differences of all terms but the leading term are zero, so the n th-order divided differences of this polynomial are constant and of value a_0 .

The result that for a polynomial of degree n , the n th-order divided differences are constant, can be used to verify whether a set of $(n+m)$

values of f can be fitted by a polynomial of the n th degree. It can also be used to determine values of this polynomial for other values of x , and so to carry out interpolation.

The latter calculation can be done by a process of building up from n th-order divided differences, rather in the way in which a polynomial of the n th order can be built up, at equal intervals in x , from its n th differences (§ 4.42). Further, it is possible to determine derivatives of this polynomial, as follows.

If $x_{j+1} = x_j + \epsilon$, then $f(x_j, x_{j+1}) = f'(x_j) + O(\epsilon)$, and in the limit $\epsilon \rightarrow 0$, $f(x_j, x_j) = f'(x_j)$. Although $f(x_j, x_j)$ cannot be evaluated directly from the values of f and the definition (5.34) of divided differences, it can be built up from *higher* orders of divided differences and so determined in this way. Similarly

$$f(x_j, x_j, x_j) = \frac{1}{2!} f''(x_j)$$

and in general
$$\underbrace{f(x_j, x_j, \dots, x_j)}_{n+1 \text{ arguments}} = \frac{1}{n!} f^{(n)}(x_j).$$

Example: To show that the following values of $f(x)$ are consistent with $f(x)$ being a cubic in x , and to find $f(6), f'(6), f''(6)$ for this cubic:

| | | | | | | |
|-----|-----|---|---|---|-----|-----|
| x | -1 | 0 | 2 | 3 | 7 | 10 |
| f | -11 | 1 | 1 | 1 | 141 | 561 |

The working can be arranged as follows:

| x | f | 1st order | 2nd order | 3rd order |
|-----|-----|---------------|-------------|-----------|
| -1 | -11 | | | |
| 0 | 1 | 12/1 = 12 | | |
| 2 | 1 | 0/2 = 0 | -12/3 = -4 | 4/4 = 1 |
| 3 | 1 | 0/1 = 0 | 0/3 = 0 | 7/7 = 1 |
| 7 | 141 | 140/4 = 35 | 35/5 = 7 | 8/8 = 1 |
| 10 | 561 | 420/3 = 140 | 105/7 = 15 | 3/3 = 1 |
| 6 | 73 | -448/-4 = 122 | -18/-1 = 18 | -1/-1 = 1 |
| 6 | | 54 | -68/-4 = 17 | -4/-4 = 1 |
| 6 | | | 13 | |
| 6 | | | | |

The working above the inclined line is concerned with showing that the third-order divided differences are constant, as is necessary for a cubic; that below the line is concerned with the evaluation of this cubic and its derivatives at $x = 6$. The arrows indicate the sequence in which the numbers in the lower part are obtained.

In the first part, the divided differences of successively *higher* orders are calculated directly from the definition; for example:

$$\begin{aligned} f(-1, 0) &= \frac{1 - (-11)}{0 - (-1)} = 12; & f(0, 2) &= \frac{1 - 1}{2 - 0} = 0; \\ f(-1, 0, 2) &= \frac{0 - 12}{2 - (-1)} = \frac{-12}{3} = -4. \end{aligned}$$

In the second part, the divided differences of successively *lower* orders are built up from those of higher orders. The value 6 of x for which $f(x)$ is wanted is written as the next value of x in the table. The value of $f(3, 7, 10, 6)$ must be the constant value 1 for this cubic, that is

$$\frac{f(7, 10, 6) - f(3, 7, 10)}{6 - 3} = 1$$

so that

$$f(7, 10, 6) - f(3, 7, 10) = 3.$$

This value is added to $f(3, 7, 10) = 15$ to give $f(7, 10, 6) = 18$. Then

$$f(7, 10, 6) = \frac{f(10, 6) - f(7, 10)}{6 - 7} = \frac{f(10, 6) - f(7, 10)}{-1}$$

so that

$$f(10, 6) - f(7, 10) = -18$$

and this is added to $f(7, 10) = 140$, to give $f(10, 6) = 122$. Finally

$$f(10, 6) = \frac{f(6) - f(10)}{6 - 10} = \frac{f(6) - f(10)}{-4}$$

so that $f(6) - f(10) = -488$, which added to $f(10)$ gives $f(6) = 73$.

This value of $f(6)$ can be checked by taking it, with the values of $f(3)$, $f(7)$, and $f(10)$, as given values of the cubic, and using them to obtain $f(2)$ in a similar way; the value obtained should reproduce the value which was used in forming the divided difference table used in the evaluation of $f(6)$.

To obtain $f'(6)$ we put a second value of 6 for x , and repeat the process as far as the first-order divided difference only, and for $\frac{1}{2}f''(6)$ we put $x = 6$ again and repeat the process as far as the second-order divided differences. The result gives the cubic in powers of $(x - 6)$; in this case

$$\begin{aligned} f &= f(6) + f'(6)(x - 6) + \frac{1}{2}f''(6)(x - 6)^2 + \frac{1}{6}f'''(6)(x - 6)^3 \\ &= 73 + 54(x - 6) + 13(x - 6)^2 + (x - 6)^3, \end{aligned}$$

and the evaluation of this for the values of x for which the function values are originally given checks the whole calculation.

5.8. Inverse interpolation

The problem of inverse interpolation is this: given a table of $f(x)$ as a function of x , to find the value of x for which $f(x)$ has a specified value. If the table is not at equal intervals in x , there is no distinction between direct and inverse interpolation; the following applies to tables at equal intervals in x , as is the case in almost all tables.

The table can be regarded as one of x at *unequal* intervals of $f(x)$, and a method of interpolation of functions given at unequal intervals of the argument (§ 5.7) can be used for inverse interpolation. This process

takes no advantage of the equal intervals in x , and needs care in use; an example of how *not* to use it is given below (§ 5.81).

Bessel’s formula to third differences, using modified second differences, is

$$f_p = f(x_0 + p \delta x) = f_0 + p \delta f_{\frac{1}{2}} + B_2(p)(\delta_m^2 f_0 + \delta_m^2 f_1) + B_3(p) \delta^3 f_{\frac{1}{2}};$$

in inverse interpolation, $f(x_0 + p \delta x)$ is given and this equation is to be solved for p . If the third-difference contribution can be neglected, it is a quadratic for p , and could be solved as such, using the conventional formula, but this is a laborious and unsuitable method for practical work; it has been said that ‘nobody but a mathematician would do it that way’.

One method is to determine p roughly by means of a graph or a few trial direct interpolations, and then make a small subtabulation in the neighbourhood of the rough value of p , at such an interval that linear interpolation can be used for the final step. This method may be found the best for occasional isolated inverse interpolations, and in the neighbourhood of turning values of $f(x)$.

Another method is to write Bessel’s formula in the form

$$p = [f_p - f_0 - B_2(p)(\delta_m^2 f_0 + \delta_m^2 f_1) - B_3(p) \delta^3 f_{\frac{1}{2}}] / \delta f_{\frac{1}{2}} \tag{5.37}$$

and use an iterative method, improved if required by the process of ‘exponential extrapolation’ (see §§ 3.4 (a) and 9.32). If second differences are modified by the use of the throwback from fourth differences, as indicated in formula (5.37), this can be used provided fourth differences are less than 500. If they are greater than this, it would be best to proceed by means of some preliminary subtabulation, for values of p in the neighbourhood of that given by (5.37).

The accuracy to which p can be determined depends on the number of figures in $\delta f_{\frac{1}{2}}$ and in assessing this accuracy it must be remembered that the last digit of $\delta f_{\frac{1}{2}}$ may be affected by rounding errors to the extent of ± 1 . Thus a value of $\delta f_{\frac{1}{2}}$ of about 200 is necessary to establish a second decimal in p .

Example: Given the following table, find $\sin^{-1} 0.4$ in degrees and decimals.

| x | $f(x) = \sin x$ | $\delta^2 f$ | $\delta^4 f$ | $-0.184 \delta^4 f$ | $\delta_m^2 f$ |
|-----|-----------------|--------------|--------------|---------------------|----------------|
| 0 | 0 | 0 | 0 | 0 | 0 |
| | 17365 | | − 528 | | |
| 10° | 0.17365 | − 528 | − 511 | 17 | − 3 − 531 |
| | 16837 | | − 480 | | |
| 20° | .34202 | − 1039 | − 435 | 31 | − 6 − 1045 |
| | 15798 | | | | |
| 30° | .50000 | − 1519 | | 45 | − 8 − 1527 |
| | 14279 | | | | |
| 40° | .64279 | − 1954 | | | |
| | 12325 | | | | |
| 50° | 0.76604 | | | | |

For interpolation in the interval $x = 20^\circ$ to 30° , we have

$$\begin{aligned}f_0 &= 34202 \\ \delta f_{\frac{1}{4}} &= 15798 \\ \delta_m^2 f_0 + \delta_m^2 f_1 &= -2572 \\ \delta^3 f_{\frac{1}{4}} &= -480\end{aligned}$$

in terms of the fifth decimal as unit. We want p for $f = 0.4$, $f - f_0 = 5798$ in terms of the fifth decimal. Hence, substituting in (5.37),

$$\begin{aligned}p &= [5798 + 2572B_2(p) + 480B_3(p)]/15798 \\ &= 0.3670_1 + 0.1628B_2(p) + 0.0304B_3(p).\end{aligned}\tag{5.38}$$

A nominal fifth decimal is kept here in $(f - f_0)/\delta f_{\frac{1}{4}}$, but not in the other terms since the quantities $B_2(p)$, $B_3(p)$ by which they are multiplied are less than $\frac{1}{10}$.

The first term in (5.38) is the value of p which would be obtained by linear interpolation. Taking it as a first approximation to p , the iterative process is as follows:

| p | r.h.s. of (5.38) |
|--------|---|
| 0.367 | $0.3670_1 + (0.1628)(-0.05808) + (0.0304)(+0.0051)$ $= 0.3670_1 - 0.0094_8 + 0.0001_5 = 0.3577_0$ |
| 0.3577 | $0.3670_1 + (0.1628)(-0.05744) + (0.0304)(+0.0054)$ $= 0.3670_1 - 0.0093_5 + 0.0001_6 = 0.3578_2.$ |

The change in the value of the right-hand side of (5.38) is only about $\frac{1}{80}$ of the change in the value of p , so the value 0.3578_2 would not be changed by more than 1 in the fifth decimal (due to rounding errors) if the right-hand side were evaluated for the better approximation $p = 0.3578_2$. The number of figures in $\delta f_{\frac{1}{4}}$ is not enough to determine the fifth decimal in p to several units. According to the purpose for which the value of $\sin^{-1} 0.4$ was wanted, it could be rounded off to four decimals, or the fifth retained as a guarding figure; if the latter course is taken it would be advisable to write it as a suffix, as a reminder that it is subject to an uncertainty of several units. Thus the result would be written $\sin^{-1} 0.4 = 23.578_2^\circ$.

For the worker who is fortunate enough to have the use of two machines simultaneously a convenient way of carrying out this successive approximation has been devised by Comrie.†

Care is necessary when carrying out inverse interpolation near a stationary value of the tabulated function. In such cases it is advisable to carry out a preliminary subtabulation so that in formula (5.37) $(\delta^2 f_0 + \delta^2 f_1)$ is not greater than $\frac{1}{2}\delta f_{\frac{1}{4}}$, before carrying out the interpolation.

5.81. How not to do inverse interpolation

The following example illustrates the dangers of trying to carry out inverse interpolation by using a Lagrange interpolation formula for x in terms of $f(x)$.‡

† See *Chambers's 6-Figure Tables*, vol. 2 (1949), Introduction, p. xxix.

‡ The warning provided by this example seems necessary, as this method has been recommended without qualification in a book on finite differences, and, moreover, in a context very similar to this example.

Example: Given the following table

| | | | | |
|---------|---|---|----|----|
| $x = 0$ | 1 | 2 | 3 | 4 |
| $y = 0$ | 1 | 8 | 27 | 64 |

find x for $y = 20$.

The five-point Lagrangian formula for x in terms of y is

$$x = y \left[\frac{(y-8)(y-27)(y-64)}{1 \cdot (-7)(-26)(-63)} \cdot 1 + \frac{(y-1)(y-27)(y-64)}{8 \cdot 7 \cdot (-19)(-56)} \cdot 2 + \right. \\ \left. + \frac{(y-1)(y-8)(y-64)}{27 \cdot 26 \cdot 19 \cdot (-37)} \cdot 3 + \frac{(y-1)(y-8)(y-27)}{64 \cdot 63 \cdot 56 \cdot 37} \cdot 4 \right] \quad (5.39)$$

and evaluation of this for $y = 20$ gives $x = -1.316$, instead of the correct value $(20)^{\frac{1}{3}} = 2.71442$. The result is not appreciably improved if one takes a six-point formula by including the value $x = 5$, $y = 125$ so that there are three points on each side of the value for which the interpolation is carried out; and a better result ($x = +2.923$) is obtained if a three-point formula involving only the values $x = 1$, 2, and 3, is used in preference to the five-point formula.

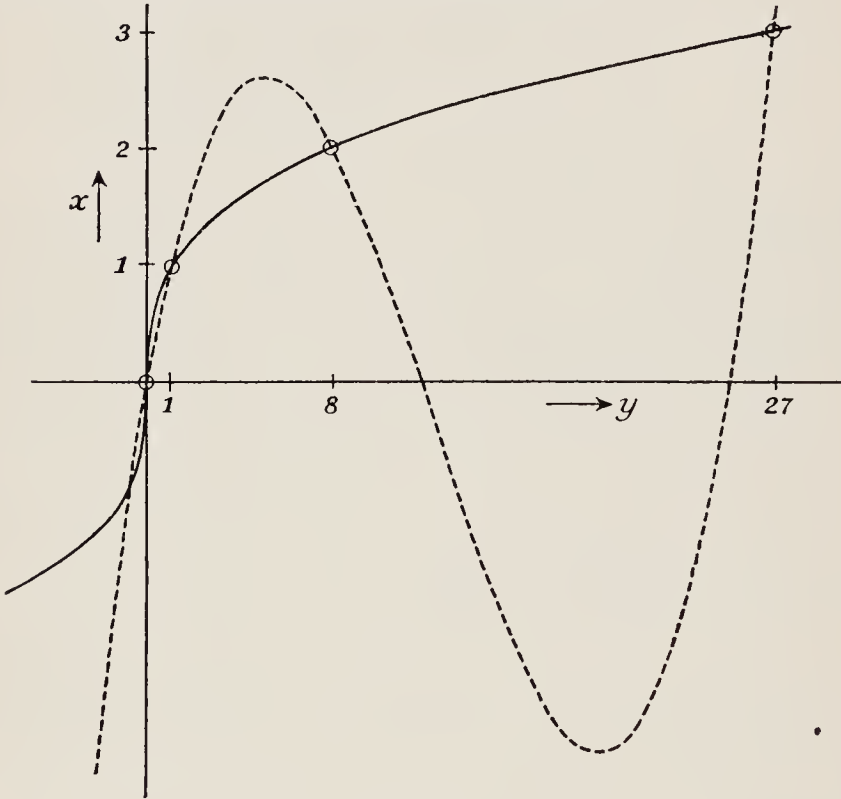


FIG. 8.

The reason for this discrepancy is that $x = y^{\frac{1}{3}}$ cannot be represented adequately by a polynomial in y over the range in question, whereas Lagrange's interpolation formula is based on a polynomial approximation to the function being interpolated. Fig. 8 shows a comparison between $x = y^{\frac{1}{3}}$ (full curve) and the quartic polynomial (5.39) by which the Lagrangian interpolation is carried out (broken curve).

If this method is used for doing inverse interpolation, it is advisable to check the resulting value of x by doing a direct interpolation in the table of $f(x)$ for that value of x , and to verify that this reproduces the value of $f(x)$.

5.9. Truncation errors in interpolation formulae

Except for polynomials of degree not greater than n , an interpolation formula to n th differences, or an $(n+1)$ -point Lagrangian interpolation formula, is only an approximation. All the formulae for direct interpolation which we have considered express the interpolated value $f(x)$ of a function as a linear combination of tabular values $f(x_j)$, or, to put it another way, as the result of some linear operation on the function specified by these tabular values. A general method for finding a formal expression for the truncation error in such a formula has been given by W. E. Milne, and is considered in the next chapter (§ 6.7).

The method of inverse interpolation considered in § 5.8 is *not* linear in f , and is not covered by Milne's treatment.

5.91. Whittaker's cardinal function

As already emphasized in § 4.1, the values f_j of a function $f(x)$ at the discrete values $x_j = x_0 + j\delta x$ of x do not define the function uniquely for intermediate values of x . However, experience of non-linear interpolation suggests that the intermediate values obtained by such an interpolation process specify a quite well-defined function, and it is of interest to inquire what this function is and just how it is related to $f(x)$. This was examined by Whittaker† and later by Ferrar;‡ this section summarizes some of their conclusions, without proofs: for proofs and fuller discussion reference should be made to the papers quoted.

Their main conclusion is this. Of all the functions $f(x)$ which have a given set of tabular values, there is one which has special properties which entitle it to be regarded in a sense as the 'simplest' function with these values, and this is the function whose values are determined by a central-difference interpolation formula regarded as an infinite series, as distinct from a truncated series which is only correct for a polynomial of finite degree. We shall consider a function $f(x)$ which has no singularities for real values of x and which tends to 0 as $|x| \rightarrow \infty$; such a function cannot be represented over the whole range of x by any polynomial approximation.

The function $\phi(y)$ defined by

$$\left. \begin{aligned} \phi(y) &= 1 && \text{for } y = 0 \\ &= (\sin \pi y)/\pi y && \text{for } y \neq 0 \end{aligned} \right\} \quad (5.40)$$

has the property that $\phi(y) = 0$ for y integral and non-zero. Hence the function $\phi[(x-x_j)/(\delta x)]f(x_j)$ has the value $f(x_j)$ at $x = x_j$ and zero at all other tabular values of x , and a function having all the tabular values $f(x_j)$ of $f(x)$ can be

† E. T. Whittaker, *Proc. Roy. Soc. Edin.* **35** (1915), 181.

‡ W. L. Ferrar, *ibid.* **45** (1925), 269; **46** (1926), 323; **47** (1927), 230.

constructed by adding a set of such contributions $\phi[(x-x_j)/(\delta x)]f(x_j)$, one for each tabular value, thus:

$$C(x) = \sum_j \phi[(x-x_j)/(\delta x)]f(x_j). \quad (5.41)$$

This function, which was introduced by Whittaker, is called the 'cardinal function' associated with $f(x)$ —or, more precisely, with the set of tabular values $f(x_j)$ of $f(x)$ at intervals δx . It is, by definition, the same function of x for all functions $f(x)$ having the same tabular values. For values of x other than tabular values,

$$\begin{aligned} \phi[(x-x_j)/\delta x] &= [\sin\{\pi(x-x_0-j\delta x)/\delta x\}]/[\pi(x-x_j)/\delta x] \\ &= (-1)^j [\sin\{\pi(x-x_0)/\delta x\}]/[\pi(x-x_j)/\delta x] \end{aligned}$$

so an alternative form for the cardinal function is

$$C(x) = (\delta x/\pi) [\sin\{\pi(x-x_0)/\delta x\}] \sum_j f(x_j)/(x-x_j).$$

From this form it is apparent that care may be necessary in summing the series for the cardinal function unless $f(x_j)$ tends to zero sufficiently fast as $|x_j| \rightarrow \infty$.

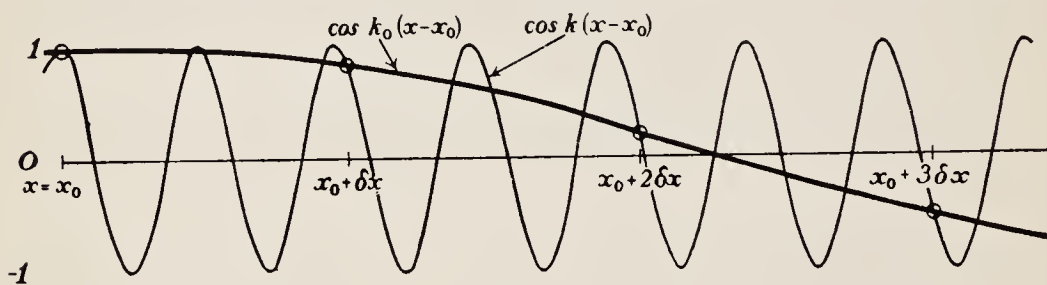


FIG. 9.

The cardinal function has three main properties, in addition to having the same tabular values as $f(x)$. First, if analysed into simple-harmonic components whose variation with x is given by $\cos[k(x-x_0)+\beta]$, then no components with periods less than $2\delta x$ (that is, with $k > \pi/\delta x$) occur. The point of this can be seen by considering $f(x)$ analysed into such components and examining what modification can be made in these components without altering the tabular values. If $A_k \cos[k(x-x_0)+\beta_k]$ is one such component, then there exists a value of n such that

$$-\pi/\delta x < k - 2n\pi/\delta x \leq \pi/\delta x;$$

if we write $k_0 = k - 2n\pi/\delta x$, then replacement of $\cos[k(x-x_0)+\beta_k]$ by

$$\cos[k_0(x-x_0)+\beta_k]$$

does not change the tabular values; this is illustrated in Fig. 9, in which the light curve represents $\cos k(x-x_0)$ and the heavy curve $\cos k_0(x-x_0)$; the ringed points indicate tabular values of x . Also $|k_0| \leq \pi/\delta x$. Now $\cos[k_0(x-x_0)+\beta_k]$ has a longer period than $\cos[k(x-x_0)+\beta_k]$; that is, the former is less rapidly oscillating function, and indeed it is the least rapidly oscillating function by which

$$\cos[k(x-x_0)+\beta_k]$$

can be replaced without affecting the tabular values. Hence if such a replacement is made for each simple-harmonic component of $f(x)$, the result is a function from which rapid oscillations have been removed as far as is possible without affecting the tabular values; and since $|k_0| \leq \pi/\delta x$, this function has no simple-harmonic components with $k > \pi/\delta x$. This is the first of the three main properties of the cardinal function, and indeed is a property which Whittaker originally used in

constructing this function. The only functions which have this property are the cardinal function itself and functions obtained by adding to it a constant multiple of $\sin[\pi(x-x_0)/\delta x]$.

A second property of the cardinal function, established by Ferrar, is one of consistency, in the following sense. Let us take the cardinal function (5.41) associated with the set of values $f(x_j)$ of $f(x)$ at intervals δx , and take another set of values of x at a different interval $\delta_1 x (> 0)$, say the values $x = x_l = x_0 + l\delta_1 x$. From the values at $x = x_l$ of the function $C(x)$ defined by (5.41), we can construct the cardinal function $C_1(x)$ associated with the values $C(x_l)$ of $C(x)$, by replacing $f(x_j)$ in (5.41) by $C(x_l)$, δx by $\delta_1 x$, and suffix j by suffix l :

$$C_1(x) = \sum_l \phi[(x-x_l)/\delta_1 x] C(x_l). \quad (5.42)$$

Now $C(x)$ defined by (5.41) is not in general the same function as $f(x)$ whose tabular values occur on the right-hand side, and correspondingly it might seem that the function $C_1(x)$ defined by (5.42) would not be the same as the function $C(x)$ whose values at interval $\delta_1 x$ appear on the right-hand side. However, Ferrar showed that provided $\delta_1 x < \delta x$, and subject to certain conditions on $f(x_j)$ for large values of $|x|$, the function $C_1(x)$ defined by (5.42) is identical with the function $C(x)$ of which $C_1(x)$ is the cardinal function. Thus for any function $f(x)$ satisfying the conditions mentioned, the function $C(x)$ defined by (5.41) has the property of reproducing itself in the operation of taking the cardinal function for any smaller interval. This property is related to that considered in the previous paragraph, for $C(x)$ has no simple-harmonic components with $k > \pi/\delta x$, and so if $\delta_1 x < \delta x$, it has none with $k > \pi/\delta_1 x$.

A third property of the cardinal function, and the one most closely related to the theory of interpolation, is that, as already mentioned, it is the function whose intermediate values are determined by a central-difference formula, considered as an infinite series, applied to the tabular values of x . The cardinal function defined by (5.41) has the same tabular values as $f(x)$ and could be used, instead of a Lagrange polynomial (5.31), for interpolation between these tabular values; it should be particularly suitable for such a purpose because of the absence of rapidly-oscillating simple-harmonic components. Whittaker showed (without use of any interpolation formula) that the expression (5.41) for $C(x)$ in terms of the tabular values of $f(x)$ could be transformed into one in terms of the differences of these tabular values; expressed in terms of Bessel's interpolation coefficients, this result is:

$$C(x_0 + p\delta x) = \sum_{n=0}^{\infty} [B_{2n}(p)\delta^{2n}f_0 + \{B_{2n}(p) + B_{2n+1}(p)\}\delta^{2n+1}f_1];$$

the right-hand side is an expression in terms of even-order differences at the beginning of the interval only, and odd-order differences, the coefficients of the latter being:

$$B_{2n}(p) + B_{2n+1}(p) = \psi_{2n+1}(p)/(2n+1)!$$

(see formulae (4.11) and (5.24)). Written symmetrically in terms of the beginning and end of the interval, this result becomes

$$C(x_0 + p\delta x) = \sum_{n=0}^{\infty} [B_{2n}(p)\{\delta^{2n}f_0 + \delta^{2n}f_1\} + B_{2n+1}(p)\delta^{2n+1}f_1]. \quad (5.43)$$

The right-hand side is the result of applying Bessel's interpolation formula (5.23), regarded as an infinite series, to the tabular values of $f(x)$ and their differences, and formula (5.43) states that the intermediate values obtained in this way are

the values of the cardinal function $C(x)$; with the conditions on $f(x)$ assumed by Whittaker, the series on the right of (5.43) always converges.

A corresponding result for the Lagrange interpolation formula with equal intervals of x has been proved by Ferrar, relating the cardinal function to the limit, as $n \rightarrow \infty$, of a symmetrical $2n$ -point Lagrange interpolation formula.

For a function $f(x) = \cos(kx + \beta)$ with $|k|\delta x < \pi$, the cardinal function $C(x)$ is identical with $f(x)$. This, and the property of the cardinal function expressed by (5.43), is one aspect of the property of the functions $\sin x$ and $\cos x$, already mentioned in § 4.1 and § 5.21, that they can be interpolated accurately from the values at a very wide interval in x , provided, of course, that enough terms are used in the interpolation formula.

VI

INTEGRATION (QUADRATURE) AND DIFFERENTIATION

6.1. Definite and indefinite integrals, and the integration of differential equations

THERE are two kinds of situation in which we may want to carry out numerical integration. One is the integration of a *given* function of the independent variable; this is sometimes called quadrature.† The other is the integration of a differential equation, which can be regarded as the evaluation of an integral in which the integrand at each value of x depends on the value of the integral at that value of x . This is represented formally by writing the solution of the equation $dy/dx = f(x, y)$ as

$$y = \int f(x, y) dx. \quad (6.1)$$

From the point of view of carrying out the integration by numerical (or mechanical) means, the only difference between quadrature and the integration of an ordinary differential equation is that in the former case the integrand in (6.1) is independent of y , and so is known as a function of the variable of integration over the whole range of x before the integration is started, whereas in the latter case the integrand at any value of x is not known until the integration has been taken as far as that value of x . The present chapter is concerned with the integration and differentiation of *given* functions of x . The integration of differential equations is considered in Chapter VII.

In integration of a given function of x , the results required may be of two kinds, a definite integral $\int_a^b f(x) dx$ between a single specified pair of limits a, b , or an indefinite integral $\int_a^x f(\xi) d\xi$ as a function of its upper limit. The latter is much the more important, and will be considered first. Usually when results of this kind are wanted, they are wanted at the same values of x as those at which $f(x)$ is tabulated, though occasionally results at twice this interval will be adequate.

† It is also sometimes called ‘mechanical quadrature’; but this term is misleading since there is nothing more mechanical about the process than there is about any other numerical calculation.

6.2. Integration formula in terms of integrand and its differences

The relation between the first differences of a function and its first derivative and the differences of this derivative has already been obtained in § 4.74, where it has been pointed out that such a relation is an integration formula. We have

$$f_1 - f_0 = \int_{x_0}^{x_1} f' dx;$$

substituting this in (4.58) and replacing f' by f we have

$$\begin{aligned} \int_{x_0}^{x_1} f dx = & \frac{1}{2}(\delta x)[f_0 + f_1 - \frac{1}{12}(\delta^2 f_0 + \delta^2 f_1) + \\ & + \frac{11}{720}(\delta^4 f_0 + \delta^4 f_1) - \frac{191}{60480}(\delta^6 f_0 + \delta^6 f_1)] + O(\delta x)^9. \end{aligned} \quad (6.2)$$

This could be obtained directly by the use of finite-difference and differential operators, without reference to § 4.74, as follows. Expressed in terms of operators, $\int_{x_0}^{x_1} f dx$ is $(E-1)D^{-1}f_0$, and we want to express this in the form $\frac{1}{2}(\delta x)\phi(\delta)(E+1)f_0$. Hence

$$\frac{1}{2}(\delta x)\phi(\delta)(E+1) = (E-1)D^{-1},$$

or
$$\phi(\delta) = \frac{E-1}{E+1} \frac{2}{U} = \frac{\tanh \frac{1}{2}U}{\frac{1}{2}U},$$

and the algebraical work of expanding $\phi(\delta)$ in powers of δ then proceeds as in § 4.74.

An alternative derivation is by integration of Bessel's or Everett's interpolation formula with respect to p , for

$$\int_{x_0}^{x_1} f dx = (\delta x) \int_{p=0}^1 f(x_0 + p\delta x) dp. \quad (6.3)$$

In Bessel's formula, the coefficients of the odd order differences are odd functions of $(p - \frac{1}{2})$, so they give zero on integration. The integrals of the coefficients of the even-order differences in Bessel's formula give the coefficients in (6.2).

The ratio of the coefficients of $\delta^4 f$ in (6.2) to that of $\delta^2 f$ is $-\frac{11}{60} = -0.1833$, which is very close to the value -0.184 used in modifying second differences in interpolation by means of the throwback; this is not surprising in view of the close relation just mentioned between the interpolation and integration formulae. However, unless modified second differences have to be calculated anyway for interpolation purposes, use of them in the integration formula is no simpler than calculating the fourth-difference contribution as it stands.

In using the integration formula (6.2) it is advisable to add the contributions in the square bracket *first* and finally multiply the whole by *one-half* (δx) , rather than dividing each separate contribution by two

before adding; this halves the possible rounding error without requiring that any additional figures should be kept.

For reference later, an alternative form of formula (6.2) should be noted. From the relation (4.34) between the operators μ and δ it follows that $(E+1)\delta^{2n} = 2(E-1)\mu\delta^{2n-1}$. So each pair of terms $(\delta^{2n}f_0 + \delta^{2n}f_1)$ can be written $2(\mu\delta^{2n-1}f_1 - \mu\delta^{2n-1}f_0)$, and formula (6.2) can be written alternatively

$$\int_{x_0}^{x_1} f dx = (\delta x) \left[\frac{1}{2}(f_0 + f_1) - \frac{1}{12}(\mu\delta f_1 - \mu\delta f_0) + \frac{11}{720}(\mu\delta^3 f_1 - \mu\delta^3 f_0) - \frac{191}{80480}(\mu\delta^5 f_1 - \mu\delta^5 f_0) \right] + O(\delta x)^9. \quad (6.4)$$

This is not as convenient as (6.2) for integration through a single interval, but may be more convenient for integration over a number of intervals.

6.21. An alternative derivation

Formula (6.2) and other integration formulae can be obtained by a rather different approach as follows. The simplest integration formula, often known as the 'trapezium rule' or the 'trapezoidal formula', is

$$\int_{x_0}^{x_1} f dx = \frac{1}{2}(\delta x)(f_0 + f_1). \quad (6.5)$$

For a more accurate formula, let us write

$$\int_{x_0}^{x_1} f dx = \frac{1}{2}(\delta x)[f_0 + f_1 + C_1]; \quad (6.6)$$

$\frac{1}{2}(\delta x)C_1$ can be regarded as a correction to the result obtained by the trapezium rule.

$$\text{Now } \int_{x_0}^{x_1} f dx = (E-1)D^{-1}f_0 \quad \text{and} \quad (f_0 + f_1) = (E+1)f_0,$$

so that C_1 is given in terms of operators by

$$C_1 = \left[\frac{2(E-1)}{U} - (E+1) \right] f_0. \quad (6.7)$$

Different integration formulae are given by different ways of expressing C_1 .

If we want to express C_1 in terms of the *sum* of contributions from the beginning and end of the interval, we write it as the result of an operation on $(E+1)f_0$, thus:

$$C_1 = \left[\frac{2(E-1)}{U(E+1)} - 1 \right] (E+1)f_0 = \left(\frac{\tanh \frac{1}{2}U}{\frac{1}{2}U} - 1 \right) (f_1 + f_0), \quad (6.8)$$

and expansion of the operator here in powers of δ gives formula (6.2). It is also possible to expand it in powers of U and so obtain an integration formula in terms of higher derivatives of f .

If we want to express C_1 in terms of the *difference* between contributions from the beginning and end of the interval, we write it as a result of an operation on $(E-1)f_0$. One way of doing this is to use in formula (6.8) the relation (4.34), namely $(E+1)\delta = 2(E-1)\mu$. This gives

$$C_1 = \frac{2}{\delta^2} \left[\frac{\tanh \frac{1}{2}U}{\frac{1}{2}U} - 1 \right] (\mu \delta f_1 - \mu \delta f_0);$$

and expansion of $(\tanh \frac{1}{2}U)/\frac{1}{2}U$ in powers of δ then gives (6.4).

6.22. Integration formula in terms of the integrand and the differences of its derivative

Another way of writing C_1 as a difference between contributions from the beginning and end of the interval is

$$C_1 = \frac{1}{D} \left[\frac{2}{U} - \frac{E+1}{E-1} \right] (E-1) Df_0 = \frac{\delta x}{U} \left[\frac{2}{U} - \frac{E+1}{E-1} \right] (f'_1 - f'_0). \quad (6.9)$$

This form for C_1 , in terms of the derivative f' of the integrand and its differences, is convenient as the operator here is an even function of U and so of δ .

From formula (4.34), $(E+1)/(E-1) = 2\mu/\delta$, so formula (6.9) for C_1 can be written

$$C_1 = -(2/\delta^2)(\delta x)[(\mu\delta/U) - (\delta^2/U^2)](f'_1 - f'_0). \quad (6.10)$$

The expansions for $(\mu\delta/U)$ and (δ^2/U^2) in terms of δ are given by (4.52) and (4.47) respectively; substitution in (6.10) gives

$$C_1 = -\frac{1}{6}(\delta x) \left[1 - \frac{1}{60}\delta^2 + \frac{1}{560}\delta^4 - \frac{79}{302400}\delta^6 \right] (f'_1 - f'_0) + O(\delta x)^9,$$

so that

$$\begin{aligned} \int_{x_0}^{x_1} f(x) dx &= \frac{1}{2}(\delta x) [(f_0 + f_1) - \frac{1}{6}(\delta x) \{ (f'_1 - f'_0) - \frac{1}{60}(\delta^2 f'_1 - \delta^2 f'_0) + \\ &\quad + \frac{1}{560}(\delta^4 f'_1 - \delta^4 f'_0) - \frac{79}{302400}(\delta^6 f'_1 - \delta^6 f'_0) \}] + O(\delta x)^{10}. \end{aligned} \quad (6.11)$$

An advantage of this formula is the small coefficient ($\frac{1}{6} \cdot \frac{1}{60} = \frac{1}{360}$) of the term of order $(\delta x)^4$ in the square bracket, compared with that ($\frac{11}{720}$) of the corresponding term in (6.2). Further, the term of order $(\delta x)^2$ in the square bracket only involves values of quantities at the beginning and end of the interval of integration, whereas the corresponding term in (6.2) involves the values of f_2 and f_{-1} outside that interval. We shall see later that both these advantages are important in the integration of differential equations.

6.23. Integration formula in terms of integrand and its derivatives (Euler–Maclaurin formula)

Expansion of the operator in (6.9) in terms of U instead of in terms of δ will give an integration formula in terms of the integrand and its derivatives. This is known as the Euler–Maclaurin formula.

In terms of U , (6.9) becomes

$$C_1 = -\{2(\delta x)/U^2\}[\frac{1}{2}U \coth \frac{1}{2}U - 1](f'_1 - f'_0). \quad (6.12)$$

Now the expansion of $\frac{1}{2}z \cot \frac{1}{2}z$ in powers of z is†

$$\frac{1}{2}z \cot \frac{1}{2}z = 1 - \frac{1}{2!}B_1 z^2 - \frac{1}{4!}B_2 z^4 - \frac{1}{6!}B_3 z^6 - \dots, \quad (6.13)$$

the coefficients B_n being the Bernoulli numbers; the values of the first few are

$$B_1 = \frac{1}{6}, \quad B_2 = \frac{1}{30}, \quad B_3 = \frac{1}{42}, \quad B_4 = \frac{1}{30}, \quad B_5 = \frac{5}{66}.$$

The corresponding expansion of $\frac{1}{2}y \coth \frac{1}{2}y$ is given by putting $z = iy$ in (6.13), so the required expansion of $(2/U^2)[\frac{1}{2}U \coth \frac{1}{2}U - 1]$ is

$$\begin{aligned} (2/U^2)[\frac{1}{2}U \coth \frac{1}{2}U - 1] &= B_1 - \frac{2}{4!}B_2 U^2 + \frac{2}{6!}B_3 U^4 - \frac{2}{8!}B_4 U^6 + O(\delta x)^8 \\ &= \frac{1}{6}[1 - \frac{1}{60}U^2 + \frac{1}{2520}U^4 - \frac{1}{100800}U^6] + O(\delta x)^8. \end{aligned} \quad (6.14)$$

Substitution in (6.12) then gives an expression for C_1 , and substitution of this in (6.6) gives the Euler–Maclaurin integration formula

$$\begin{aligned} \int_{x_0}^{x_1} f(x) dx &= \frac{1}{2}(\delta x)[f_0 + f_1 - \frac{1}{6}(\delta x)\{(f'_1 - f'_0) - \frac{1}{60}(\delta x)^2(f'''_1 - f'''_0) + \\ &\quad + \frac{1}{2520}(\delta x)^4(f^{v}_1 - f^{v}_0) - \frac{1}{100800}(\delta x)^6(f^{vii}_1 - f^{vii}_0)\}] + O(\delta x)^9. \end{aligned} \quad (6.15)$$

This formula is of limited practical value, since values of the higher derivatives of the integrand will not generally be available. They may, however, be available in two cases; first, when the integrand is given by a sufficiently simple analytical formula, and secondly, when f satisfies a sufficiently simple differential equation. The analytical formula or differential equation must be such that it can be differentiated several times without leading to expressions too complicated for practical numerical evaluation.

6.3. Integration over more than one interval

If it is adequate to obtain the values of the integral at intervals greater than those at which the integrand is given, other integration

† E. T. Whittaker and G. N. Watson, *Modern Analysis* (C.U.P. 1927), § 7.2. The notation for the Bernoulli numbers used here follows that of Whittaker and Watson.

formulae are available. For integration over $2k$ successive intervals (δx) we have

$$\int_{x-k}^{x_k} f(x) dx = (E^k - E^{-k})D^{-1}f_0 = 2(\sinh kU)D^{-1}f_0.$$

A first approximation is $2k(\delta x)f_0$, so we try to obtain a $\phi(\delta)$ such that

$$\int_{x-k}^{x_k} f(x) dx = 2k(\delta x)\phi(\delta)f_0.$$

The operator $\phi(\delta)$ required is therefore given by

$$2k(\delta x)\phi(\delta) = 2(\sinh kU)/D,$$

$$\text{or} \quad \phi(\delta) = (\sinh kU)/kU. \quad (6.16)$$

In particular, for $k = 1$ (integration over two intervals),

$$\phi(\delta) = (\sinh U)/U.$$

The expansion of this operator in powers of δ has already been considered in § 4.75; substitution from formula (4.52) gives

$$\int_{x-1}^{x_1} f(x) dx = 2(\delta x)[f_0 + \frac{1}{6}\delta^2 f_0 - \frac{1}{180}\delta^4 f_0 + \frac{1}{1512}\delta^6 f_0] + O(\delta x)^9. \quad (6.17)$$

The first two terms give the finite-difference form of the integration formula usually called 'Simpson's rule'. This can be seen by expressing $\delta^2 f_0$ in terms of function values; then the first two terms in (6.17) give

$$\int_{x-1}^{x_1} f(x) dx = 2(\delta x)[f_0 + \frac{1}{6}(f_1 - 2f_0 + f_{-1})] = \frac{1}{3}(\delta x)[f_1 + 4f_0 + f_{-1}], \quad (6.18)$$

the usual form of Simpson's rule.

Another important formula of this kind is related to the result of putting $k = 3$ in (6.16). For $k = 3$,

$$\begin{aligned} \phi(\delta) &= (\sinh 3U)/3U = 1 + \frac{3}{2}U^2 + \frac{27}{40}U^4 + \frac{81}{560}U^6 + O(\delta x)^8 \\ &= [1 + \frac{3}{2}\delta^2 + \frac{11}{20}\delta^4 + \frac{41}{840}\delta^6] + O(\delta x)^8. \end{aligned}$$

If now we replace the coefficient $\frac{41}{840}$ by $\frac{42}{840} = \frac{1}{20}$ we obtain a formula which is certainly not correct to sixth differences of the integrand but in which almost the whole of the contribution from the sixth difference is included, namely

$$\int_{x-3}^{x_3} f(x) dx = \frac{3}{16}(\delta x)[20f_0 + 30\delta^2 f_0 + 11\delta^4 f_0 + \delta^6 f_0] - \frac{1}{140}(\delta x)\delta^6 f_0 + O(\delta x)^9,$$

or, in terms of function values, taking only the terms in the square brackets,

$$\int_{x_{-1}}^{x_3} f(x) dx = \frac{3}{10}(\delta x)[f_3 + 5f_2 + f_1 + 6f_0 + f_{-1} + 5f_{-2} + f_{-3}]. \quad (6.19)$$

This is known as 'Weddle's rule'.

Another procedure for evaluating an integral over a number of equal intervals is to express it as the sum of a number of trapezium-rule contributions and a correction.

The approximation to $\int_{x_0}^{x_n} f(x) dx$ as the sum of a number of trapezium-rule contributions over intervals δx is

$$\begin{aligned} \frac{1}{2}(\delta x)(f_0 + 2f_1 + 2f_2 + \dots + 2f_{n-1} + f_n) \\ = \frac{1}{2}(\delta x)(1 + E)(1 + E + E^2 + \dots + E^{n-1})f_0 \\ = \frac{1}{2}(\delta x)[(E + 1)(E^n - 1)/(E - 1)]f_0. \end{aligned}$$

The integral itself is $(E^n - 1)D^{-1}f_0$; let us write it as the sum of the trapezium-rule contributions, plus a correction $\frac{1}{2}(\delta x)C_n$; that is,

$$(E^n - 1)D^{-1}f_0 = \int_{x_0}^{x_n} f(x) dx = \frac{1}{2}(\delta x) \left[\frac{E + 1}{E - 1} (E^n - 1)f_0 + C_n \right].$$

Then
$$C_n = \left(\frac{2}{U} - \frac{E + 1}{E - 1} \right) (E^n - 1)f_0 = \left(\frac{2}{U} - \frac{E + 1}{E - 1} \right) (f_n - f_0),$$

so that C_n is related to $(f_n - f_0)$ in just the same way as C_1 is to $(f_1 - f_0)$ (see § 6.21). Thus we can write down three integration formulae directly from the results of §§ 6.2 to 6.23.

In terms of the integrand f and its differences:

$$\begin{aligned} \int_{x_0}^{x_n} f(x) dx = (\delta x) \left[\frac{1}{2}(f_0 + 2f_1 + 2f_2 + \dots + 2f_{n-1} + f_n) - \frac{1}{12}(\mu\delta f_n - \mu\delta f_0) + \right. \\ \left. + \frac{11}{720}(\mu\delta^3 f_n - \mu\delta^3 f_0) - \frac{191}{60480}(\mu\delta^5 f_n - \mu\delta^5 f_0) \right]. \quad (6.20) \end{aligned}$$

In terms of the integrand, its first derivative, and the differences of this first derivative:

$$\begin{aligned} \int_{x_0}^{x_n} f(x) dx = \frac{1}{2}(\delta x)[f_0 + 2f_1 + 2f_2 + \dots + 2f_{n-1} + f_n - \\ - \frac{1}{6}(\delta x)\{(f'_n - f'_0) - \frac{1}{60}(\delta^2 f'_n - \delta^2 f'_0) + \frac{1}{6060}(\delta^4 f'_n - \delta^4 f'_0) - \dots\}]. \quad (6.21) \end{aligned}$$

In terms of the integrand and its derivatives (Euler–Maclaurin formula):

$$\int_{x_0}^{x_n} f(x) dx = \frac{1}{2}(\delta x)[f_0 + 2f_1 + 2f_2 + \dots + 2f_{n-1} + f_n - \frac{1}{6}(\delta x)\{(f'_n - f'_0) - \frac{1}{60}(\delta x)^2(f'''_n - f'''_0) + \frac{1}{2520}(\delta x)^4(f^{(5)}_n - f^{(5)}_0) - \dots\}]. \quad (6.22)$$

The last is of limited practical use as a formula for numerical integration for reasons already mentioned in § 6.23. It is, however, useful for numerical work in another context (see § 11.12).

6.4. Evaluation of an integral as a function of its upper limit

The evaluation of an integral as a function of the upper limit can be carried out by successive addition of the contributions from a sequence of intervals (δx) covering the relevant range of x . Let us for brevity write $F(x)$ for $\int_a^x f(\xi) d\xi$. Then if, for example, the integration formula (6.2) is used, these contributions are

$$\delta F_{j+\frac{1}{2}} = \int_{x_j}^{x_{j+1}} f(x) dx = \frac{1}{2}(\delta x)s_{j+\frac{1}{2}}, \quad (6.23)$$

$$\text{where } s_{j+\frac{1}{2}} = (f_j + f_{j+1}) - \frac{1}{12}(\delta^2 f_j + \delta^2 f_{j+1}) + \frac{11}{720}(\delta^4 f_j + \delta^4 f_{j+1}) - \dots; \quad (6.24)$$

and $\int_{x_0}^{x_n} f(x) dx$ is the sum of n such contributions. The summation of these contributions can be expressed by use of the central-sum operator σ , the inverse of δ (see § 4.6); operating on both sides of (6.23) with σ we have

$$F_j - F_0 = \frac{1}{2}(\delta x)[(\sigma s)_j - (\sigma s)_0]. \quad (6.25)$$

In summing contributions of the form (6.23), there will be an accumulation of rounding errors from the corrections to the trapezium rule for the successive intervals. This accumulation can be made unimportant by the use of a guarding figure in the contributions (6.23).

In whatever way the calculation of the integral is done, it must be checked. The details of the checking procedure depend on the method used to evaluate the integral; the following procedure, given as an example, refers to a calculation carried out by evaluating the successive contributions (6.23) and adding them. It is then advisable to carry out one check on the values of $s_{j+\frac{1}{2}}$ and another on the evaluation of the integral from them; the intermediate check of the values of $s_{j+\frac{1}{2}}$ will avoid the possibility of a large number of values of the integral having to be corrected if one of the values of $s_{j+\frac{1}{2}}$ is in error.

Operating on both sides of (6.24) with δ^2 we have

$$\delta^2 s_{j+\frac{1}{2}} = (\delta^2 f_j + \delta^2 f_{j+1}) - \frac{1}{12}(\delta^4 f_j + \delta^4 f_{j+1}) + \frac{11}{720}(\delta^6 f_j + \delta^6 f_{j+1}) \dots \quad (6.26)$$

Values of $\delta^2 s$ can be calculated (i) from successive values of s by the method of § 4.45, and (ii) from formula (6.26); comparison of these values provides a good check on the s values, and a clear indication of the location of mistakes, if any. When this check has been made and mistakes, if any, corrected, and only then, the values of $s_{j+\frac{1}{2}}$ should be summed, and the values of the integral F calculated from (6.25); the multiplication by $\frac{1}{2}\delta x$ should follow the summation, otherwise it may be necessary to keep an extra figure to avoid rounding errors.

This summation and multiplication by $\frac{1}{2}\delta x$ can be checked as follows. From (6.23) we have

$$\delta^3 F_{j+\frac{1}{2}} = \frac{1}{2}(\delta x)\delta^2 s_{j+\frac{1}{2}}, \quad (6.27)$$

and values of $\delta^2 s$ are already available as they have been used in the process of checking the values of s . The values of the integral F can be checked by comparing the values of $\delta^3 F$ obtained (i) as $\delta(\delta^2 F)$ from the values of F , the second differences being evaluated by the method of § 4.45 and *then* differenced once more, and (ii) from formula (6.27).

Various alternative checking procedures of this kind can be devised, for example by taking the fourth differences of both sides of (6.24) to give a check of the values of s . Another kind of check is an overall check by one of the methods considered in §§ 6.5, 6.6; but this is less useful, since although it will indicate the presence of a mistake it will not usually locate it.

Example: To evaluate $\int_0^1 e^{x^2} dx$ to five decimals at intervals 0.1 in x .

The first seven intervals of the calculation are given on p. 107, in which each column, rather than each row, refers to a single value of x . The values of s are calculated from formulae (6.24), the contributions from the sixth differences being just appreciable; the second differences of these values of s are in the line below the values of s themselves, and the sums of the right-hand side of formula (6.26) occur four lines lower down. The discrepancies (+60, -121, +61) between these two sets of values at $x = 0.35, 0.45$, and 0.55 clearly indicate a mistake, probably of 60 units in the last figure, at $x = 0.45$, and this is easily traced to the value of $-\frac{1}{12}(\delta^2 f_j + \delta^2 f_{j+1})$ for the interval 0.4 to 0.5, which should read -583_0 , so that $s(0.45) = 2.45180$.

The differences $\delta^2 s$ affected by this correction should be recalculated and the comparison with formula (6.26) verified to make sure that the correction itself has been rightly made. In actual working the corrected values could be written in place of the erroneous values, but in this example they have been written separately to display both the incorrect and correct values. The successive values of s are then summed, and the sums multiplied by $\frac{1}{2}\delta x$ and rounded off. Finally, this summation and multiplication is checked by comparing $\delta(\delta^2 F)$ with $\frac{1}{2}(\delta x)\delta^2 s$.

Notes: (i) In the check comparisons between the values of δ^2s calculated by the two ways, and in the similar comparisons for δ^3F , discrepancies of a unit can be expected as the result of accumulated rounding errors, but discrepancies in successive values should usually be in opposite directions.

(ii) The contribution of δ^6f to δ^2s may be appreciable, although its contribution to s itself is negligible, and similarly for other orders of differences.

(iii) In this case f is an even function of x , hence each difference of even order is an even function of x , so the even-order differences at $x = 0$ can be calculated from the formula

$$\delta^{2n+2}f(0) = 2[\delta^{2n}f(\delta x) - \delta^{2n}f(0)].$$

To obtain the second and higher differences at $x = 0.8$, the values of $f(x)$ up to $f(1.1)$ have been used, though to save space they are not given explicitly.

(iv) Each value of f is subject to a rounding error of up to $\frac{1}{2}$ in the last figure, so each value of s is subject to a rounding error of up to 1. If these rounding errors were randomly distributed, the 'probable error' (in the technical sense of the term as used in the theory of errors) of the sum of N values of s would be approximately $\frac{1}{2}N^{\frac{1}{2}}$, so that of the result would be about $\frac{1}{4}N^{\frac{1}{2}}(\delta x)$ in the last significant figure of the values of the integrand. In the present case, with $N = 8$, this is rather less than 1 in the sixth decimal in the integral. In making estimates of this kind, it must be remembered that errors up to 2 or 3 times the 'probable error' are not unlikely.

(v) Another method of calculating this integral, more convenient for large values of x , is given in the example in § 7.3.

(vi) The checks indicated verify the calculation from the values of the integrand f but they do not check these values themselves. The differences of the values of the integrand form a check against gross errors, but do not check the last digit with certainty. An overall check on the whole calculation, including the values of the integrand, is provided by doing an integration by one of the methods of § 6.5 using another set of values of the integrand, say in this case by integrating from $x = 0$ to 0.8 using ten steps of length 0.08, or by using a Gauss integration formula (see § 6.61).

The following is a convenient alternative procedure if values of the integrand are available at an interval δx which is *half* that at which the integral is required. From formula (6.17), neglecting terms $O(\delta x)^9$,

$$\int_{x_{2j}}^{x_{2j+2}} f(x) dx = 2(\delta x) \left[f_{2j+1} + \frac{1}{6} \delta^2 f_{2j+1} - \frac{1}{180} \delta^4 f_{2j+1} + \frac{1}{1512} \delta^6 f_{2j+1} \right]$$

so that for integration over a range x_0 to x_{2J} ,

$$\int_{x_0}^{x_{2J}} f(x) dx = 2(\delta x) \left[\sum_{j=0}^{J-1} f_{2j+1} + \frac{1}{6} \sum_{j=0}^{J-1} \delta^2 f_{2j+1} - \frac{1}{180} \sum_{j=0}^{J-1} \delta^4 f_{2j+1} + \frac{1}{1512} \sum_{j=0}^{J-1} \delta^6 f_{2j+1} \right] \quad (6.28)$$

the sums being over *alternate* values of the integrand and of its differences of even order. This formula, taken as far as the term in $\delta^2 f$, is probably the most convenient form of Simpson's rule for practical work, and the higher-difference terms can easily be included if appreciable.

| x | -0.1 | 0 | +0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 |
|---|---------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|--------------------|---------|
| $f = e^{x^2}$ | 1.01005 | 1.00000 | 1.01005 | 1.04081 | 1.09417 | 1.17351 | 1.28403 | 1.43333 | 1.63232 | 1.89648 |
| $\delta^2 f$ | | 2010 | 2071 | 2260 | 2598 | 3118 | 3878 | 4969 | 6517 | 8727 |
| $\delta^4 f$ | | 122 | 128 | 149 | 182 | 240 | 331 | 457 | 662 | 957 |
| $\delta^6 f$ | | 12 | 15 | 12 | 25 | 33 | 35 | 79 | 90 | 170 |
| $f_j + f_{j+1}$ | | 2.01005 | 2.05086 | 2.13498 | 2.26768 | 2.45754 | 2.71736 | 3.06565 | 3.52880 | |
| $-\frac{1}{12}(\delta^2 f_j + \delta^2 f_{j+1})$ | | -340 ₁ | -360 ₉ | -404 ₈ | -476 ₃ | -523 ₀ | -737 ₂ | -957 ₃ | -1270 ₃ | |
| $+\frac{1}{720}(\delta^4 f_j + \delta^4 f_{j+1})$ | | +3 ₈ | +4 ₂ | +5 ₁ | +6 ₄ | +8 ₇ | +12 ₀ | +17 ₁ | +24 ₈ | |
| $-\frac{1}{60480}(\delta^6 f_j + \delta^6 f_{j+1})$ | | -0 ₁ | -0 ₁ | -0 ₁ | -0 ₂ | -0 ₂ | -0 ₄ | -0 ₅ | -0 ₈ | |
| sum = $s_{j+\frac{1}{2}}$ | 2.00669 | 2.00669 | 2.04729 | 2.13098 | 2.26298 | 2.45240 | 2.71010 | 3.05624 | 3.51634 | |
| $\delta^2 s$ | | 4060 | 4309 | 4831 | 5742 | 6828 | 8844 | 11396 | | |
| $\delta^2 f_j + \delta^2 f_{j+1}$ | | 4081 | 4331 | 4858 | 5716 | 6996 | 8847 | 11486 | | |
| $-\frac{1}{12}(\delta^4 f_j + \delta^4 f_{j+1})$ | | -20 ₈ | -23 ₁ | -27 ₆ | -35 ₂ | -47 ₈ | -65 ₇ | -93 ₂ | | |
| $+\frac{1}{720}(\delta^6 f_j + \delta^6 f_{j+1})$ | | +0 ₄ | +0 ₄ | +0 ₈ | +0 ₉ | +1 ₀ | +1 ₇ | +2 ₆ | | |
| sum | | 4061 | 4308 | 4831 | 5682 | 6949 | 8783 | 11395 | | |
| corrected s | 2.00669 | 2.04729 | 2.13098 | 2.26298 | 2.45180 | 2.71010 | 3.05624 | 3.51634 | | |
| $\delta^2 s$ | | 4060 | 4309 | 4831 | 5682 | 6948 | 8784 | 11396 | | |
| os | 0.00000 | 2.00669 | 4.05398 | 6.18496 | 8.44794 | 10.89974 | 13.60984 | 16.66608 | 20.18242 | |
| $\frac{1}{2}\delta x os = \int_0^x f dx = F$ | 0.00000 | 0.10033 | 0.20270 | 0.30925 | 0.42240 | 0.54499 | 0.68049 | 0.83330 | 1.00912 | |
| $\delta^2 F$ | 0 | 204 | 418 | 660 | 944 | 1291 | 1731 | 2301 | | |
| $\delta(\delta^2 F)$ | | 204 | 214 | 242 | 284 | 347 | 440 | 570 | | |
| $\frac{1}{2}(\delta x)\delta^2 s$ | | 203 ₀ | 215 ₄ | 241 ₅ | 284 ₁ | 347 ₄ | 439 ₂ | 569 ₈ | | |

Example: To evaluate $\int_0^1 e^{x^2} dx$ to five decimals at intervals 0.2 in x , using values of the integrand at intervals $\delta x = 0.1$.

The integration to $x = 0.8$ is given on p. 109. The integrand values and their differences are the same as on p. 107, but now the values of x are arranged in a column instead of in a line. The values of the integrand and its differences which are to be used are enclosed in 'boxes'; this is a convenient way of picking out these values and ensuring that the right ones are used. The columns headed $\sum f$ and $\sum \delta^{2n}f$ contain current sums of the 'boxed' values of f and its differences of even order, and are placed on the lines corresponding to the values of x for which they will be used in evaluating the integral by means of formula (6.28).

The results can be checked by a formula, similar to (6.27), obtained by taking third differences, at the large interval, of both sides of formula (6.28), or by one of the methods of §§ 6.5, 6.6.

6.41. Change of interval length in an integration

It may sometimes be required to change the interval length (δx) in the course of an integration. Where the third and higher derivatives of the integrand are large, it is advisable to take small intervals, not only in order that the corrections to the trapezium rule should not be too large, but in order that the behaviour of the integrand should be adequately defined by those of its values which are used in the integration formula. It may happen that the values of the third and higher derivatives vary considerably over the range of integration. If in such a case the interval length (δx) which is necessary when these derivatives are large were used over the whole range, this would involve an unnecessary amount of work in the region in which they are small. So a change of interval length, or several such changes, may be required in the course of the integration.

When such a change is made, there should *always* be an overlap between the ranges of x for which the different sizes of interval are used. This is a potent check against mistakes, which are particularly likely to be made at points like this at which a systematic procedure is interrupted. It provides a check not only against random mistakes but against some forms of systematic mistakes as well; if, for example, the term $-\frac{1}{12}(\delta^2 f_j + \delta^2 f_{j+1})$ in (6.24) had been taken systematically with the wrong sign, this would be shown up by a discrepancy between the integration carried out with two different interval lengths.

In practice, the convenient intervals are 1, 2, and 5 times a power (positive or negative) of 10; an increase of interval length from 2.10^q to 5.10^q involves some interpolation, but this is all of the simplest kind, namely 'halfway' interpolation (see § 5.21), and should give no trouble.

| x | $f = e^{x^2}$ | $\delta^2 f$ | $\delta^4 f$ | $\delta^6 f$ | $\sum \delta^2 f$ 0 | $\sum \delta^4 f$ 0 | $\sum \delta^6 f$ 0 | $\sum f + \frac{1}{6} \sum \delta^2 f - \frac{1}{180} \sum \delta^4 f = S$ 0 | $2(\delta x)S$ 0 |
|-----|----------------|--------------|--------------|--------------|------------------------|------------------------|------------------------|---|---------------------|
| 0.0 | 1.00000 | 2010 | 122 | 12 | | | | | |
| 0.1 | <u>1.01005</u> | <u>2071</u> | <u>128</u> | <u>15</u> | | | | | |
| 0.2 | 1.04081 | 2260 | 149 | 12 | 2071 | 128 | 15 | 1.01005 + 345 - 1 = 1.01349 | 0.20270 |
| 0.3 | <u>1.09417</u> | <u>2598</u> | <u>182</u> | <u>25</u> | | | | | |
| 0.4 | 1.17351 | 3118 | 240 | 33 | 4669 | 310 | 40 | 2.10422 + 778 - 2 = 2.11198 | 0.42240 |
| 0.5 | <u>1.28403</u> | <u>3878</u> | <u>331</u> | <u>35</u> | | | | | |
| 0.6 | 1.43333 | 4869 | 457 | 79 | 8547 | 641 | 75 | 3.38825 + 1425 - 4 = 3.40246 | 0.68049 |
| 0.7 | <u>1.63232</u> | <u>6517</u> | <u>662</u> | <u>90</u> | | | | | |
| 0.8 | 1.89648 | 8727 | 957 | 170 | 15064 | 1303 | 165 | 5.02057 + 2511 - 7 = 5.04561 | 1.00912 |

6.42. Integration in the neighbourhood of a singularity of the integrand

A point at which f or any of its derivatives becomes infinite will be called a 'singularity' of f .

In deriving the integration formulae of the previous sections, it has been assumed that the integrand $f(x)$ is expansible in a Taylor series through each interval δx . This is not the case if there is a singularity at any point (including end-points) of the interval, and the approximations used are likely to be bad in the neighbourhood of a singularity even if it does not lie in the interval through which the integration is being taken. Examples are:

$$(i) \int_1 \frac{e^{-ax}}{(x^2-1)^{\frac{1}{2}}} dx \quad (\text{integrand infinite at } x = 1);$$

$$(ii) \int_0 x^{\frac{1}{2}} \sin x \, dx \quad (\text{second derivative of integrand infinite at } x = 0).$$

A singularity can often be removed by a change of independent variable; for example, the substitution $x = \cosh u$ makes

$$\int_1 \frac{e^{-ax}}{(x^2-1)^{\frac{1}{2}}} dx = \int_0 e^{-a \cosh u} du,$$

and usually such a change of independent variable will also give results in a more satisfactory form for interpolation. But if the results are required in terms of x , and in a context in which no interpolation will be carried out on them, it may sometimes be better to obtain them directly in such a form.

A singularity can sometimes be removed by the following process. We subtract from the integrand f a function g which can be integrated formally and has a singularity of the same kind as f , evaluate the analytical formula for $\int g \, dx$, and evaluate $\int (f-g) \, dx$ by numerical integration. This may be called 'subtracting out the singularity'.

This is satisfactory if the singularity is a pole of order n , but otherwise it may not be possible to remove the singularity in this way, though it may be made less severe. For example, we can write

$$\int_1 \frac{e^{-ax}}{(x^2-1)^{\frac{1}{2}}} dx = e^{-a} \int_1 \frac{dx}{(x^2-1)^{\frac{1}{2}}} + \int_1 \frac{e^{-ax} - e^{-a}}{(x^2-1)^{\frac{1}{2}}} dx.$$

The integrand in the integral on the left is infinite at $x = 1$, whereas that in the second integral on the right is finite, though its derivative is infinite.

An alternative, and often more effective, treatment is as follows. Write the integral $\int f(x) dx = g(x)h(x)$, where g is a chosen function for which dg/dx has a singularity of the same kind as that of f . This leads to a differential equation for $h(x)$ for which, however, numerical integration may be quite practicable. This may be called 'dividing out the singularity'.

Consider, for example, integrals of the form

$$\int_0 x^p f(x) dx,$$

where $f(x)$ is regular at $x = 0$ and $f(0) \neq 0$, and p is greater than -1 and is not an integer. For this case the appropriate function $g(x)$ is $x^{p+1}/(p+1)$, so we write

$$\int_0 x^p f(x) dx = \frac{1}{p+1} x^{p+1} h(x) \quad (6.29)$$

and on differentiation obtain

$$x \frac{dh}{dx} = (p+1)(f-h). \quad (6.30)$$

On differentiating k times and putting $x = 0$, this gives

$$kh^{(k)}(0) = (p+1)[f^{(k)}(0) - h^{(k)}(0)],$$

so that

$$h^{(k)}(0) = \frac{p+1}{p+1+k} f^{(k)}(0). \quad (6.31)$$

In particular

$$\left. \begin{aligned} h(0) &= f(0) \\ h'(0) &= \frac{p+1}{p+2} f'(0) \\ h''(0) &= \frac{p+1}{p+3} f''(0) \end{aligned} \right\}. \quad (6.32)$$

These serve as starting values for a numerical integration of equation (6.30).

Numerical integration of equation (6.30) is also useful as a means of evaluating integrals of the form $\int x^p f(x) dx$ for values of the lower limit in the neighbourhood of zero when p is negative and $|p| > 1$.

6.43. Integration when the integrand increases 'exponentially'

A similar device of writing an integral in the form

$$\int f(x) dx = g(x)h(x),$$

choosing a convenient function $g(x)$, and solving numerically the result-

ing differential equation for $h(x)$ can often be applied when the integrand increases rapidly with x , particularly when $\log f(x)$ increases more rapidly than linearly in x . This process then consists of dividing out the singularity at infinity. One good choice of $g(x)$ is the leading term in the asymptotic behaviour of $\int f(x) dx$; this makes $h \rightarrow 1$ as $x \rightarrow \infty$. For example, $\int_0^x e^{x^2} dx$ behaves asymptotically like $e^{x^2}/2x$, so we may write

$$\int_0^x e^{x^2} dx = (e^{x^2}/2x)h(x), \quad (6.33)$$

then
$$\frac{dh}{dx} + 2x \left[\left(1 - \frac{1}{2x^2}\right)h - 1 \right] = 0. \quad (6.34)$$

This, however, is clearly not convenient for small x , and in order to obtain an equation applicable to the whole range of x we may be content to divide out the main part of the singularity at infinity by writing

$$\int_0^x e^{x^2} dx = e^{x^2}h,$$

then
$$\frac{dh}{dx} + 2xh = 1, \quad (6.35)$$

a simpler equation than (6.34), and one which there is no difficulty in integrating numerically from $x = 0$ (see the example in § 7.3).

6.44. Twofold integration

By a many-fold integration of a function $f(x)$ is meant a result obtained by repeated integration with respect to the *same* independent variable, as distinct from a double integral which is obtained by one integration with respect to each of two independent variables. That is, a twofold integral of a function f is a function F such that

$$\frac{d^2 F}{dx^2} = f(x).$$

It is sometimes convenient to be able to obtain such a twofold integral directly without going through the intermediate stage of evaluating $\int f dx$. This can be done by using formula (4.54), which for this case becomes

$$\delta^2 F_0 = (\delta x)^2 [f_0 + \frac{1}{12}\delta^2 f_0 - \frac{1}{240}\delta^4 f_0 + \frac{31}{80480}\delta^6 f_0] + O(\delta x)^{10}. \quad (6.36)$$

The twofold summation of the second differences $\delta^2 F$ to give the function values F can be done either directly by the method of § 4.46 or by obtaining the first differences as an intermediate step and then summing these.

Effects of rounding errors may build up somewhat rapidly in this twofold summation, and it is advisable to carry some guarding figures.

The twofold integration could be carried out by two single integrations, one from F'' to F' and the other from F' to F . Suppose that in an integration carried out by this method, n decimals would have been kept in F' . Then in a twofold integration carried out by a method not involving the calculation of F' , the number of decimals kept in δF (or in F if δF is not calculated) should be enough to give n decimals in $\delta F/\delta x$.

6.5. Integrals between fixed limits

An integral between fixed limits can be evaluated by any of the formulae of § 6.3 or § 6.4, the difference being that the value of the integral is only wanted for a single value of the upper limit.

Example: To evaluate $\int_0^{0.8} e^{x^2} dx$ using formula (6.21).

In this case $f'(x) = 2xe^{x^2}$ is an odd function of x , hence all even-order differences of $f'(x)$ at $x = 0$ are zero, and their contribution to formula (6.21) is zero. For the even-order differences at $x = 0.8$ we have the following values:

| x | $f'(x) = 2xe^{x^2}$ | $\delta^2 f'$ | $\delta^4 f'$ | $\delta^6 f'$ | |
|-----|---------------------|---------------|---------------|---------------|------------------------------------|
| 0.5 | 1.2840 | | | | At $x = 0.8$ |
| .6 | 1.7200 | 1292 | | | $f' = 3.0344$ |
| .7 | 2.2852 | 1840 | 238 | | $-\frac{1}{80}\delta^2 f' = -43_8$ |
| .8 | 3.0344 | 2626 | 374 | 50 | $+\frac{1}{560}\delta^4 f' = +0_7$ |
| .9 | 4.0462 | 3786 | 560 | | <u>Sum = 3.0301</u> |
| 1.0 | 5.4366 | 5506 | | | |
| 1.1 | 7.3776 | | | | |

The values of $f(x)$ are given in the example in § 6.4 (p. 107); using them we have

$$2 \sum_{j=1}^7 f(j\delta x) = 17.33644$$

$$f(0) + f(0.8) = 2.89648$$

and, from the values of f' above,

$$-\frac{1}{8}(\delta x)(f' - \frac{1}{60}\delta^2 f' + \frac{1}{560}\delta^4 f') = \frac{-0.05050}{20.18242}$$

Hence
$$\int_0^{0.8} e^{x^2} dx = \frac{1}{2}(0.1)20.18242 = 1.00912_1.$$

The tolerance on this result, due to the accumulation of rounding errors of the function values used, is a few units in the sixth decimal.

Some other forms which are only appropriate to an integral between fixed limits are considered in the following sections. In particular, there is the possibility of using values of the integrand $f(x)$ not spaced at equal intervals in x if there is any advantage in doing so (§ 6.6). If the integrand is specified by a table at equal intervals in x , then an integration formula which makes use of this feature is usually the more

convenient; the interpolation necessary to give its values for use in an integration formula using unequally spaced values of x would usually outweigh the advantages of such a formula. But if it is specified by a formula which can equally well be evaluated for any value of x , then use of values of x at unequal intervals may become practicable.

6.51. Gregory's formula

The integration formula (6.20) expresses the correction C_n to the sum of a set of trapezium rule contributions in terms of central differences at the beginning and end of the range of integration. If the only available values of the integrand are those from f_0 to f_n , the ends of the range of integration, then only forward differences from the beginning and backward differences from the end of the range are available, and a formula in terms of these differences is needed. This is

$$\int_{x_0}^{x_n} f(x) dx = (\delta x) \left[\frac{1}{2}(f_0 + 2f_1 + 2f_2 + \dots + 2f_{n-1} + f_n) + \left(\frac{1}{12}\Delta - \frac{1}{24}\Delta^2 + \frac{1}{720}\Delta^3 - \dots \right) f_0 - \left(\frac{1}{12}\nabla + \frac{1}{24}\nabla^2 + \frac{1}{720}\nabla^3 + \dots \right) f_n \right] \quad (6.37)$$

and is known as *Gregory's formula*.

Its practical use is limited, because usually the reason for the limitation on the range over which a function is defined is the occurrence of a singularity, as at $x = \pm 1$ in $\int_{-1}^1 e^{-x^2}(1-x^2)^{-\frac{1}{2}} dx$, in which case the extension of the integration formula up to those points is invalid. In most other cases, values of the integrand outside the range of integration, and therefore the central differences required in formula (6.20), are available, and then this formula should *always* be used in preference to Gregory's formula.†

6.52. Integral in terms of function values

By expressing the differences in formula (6.20), (6.21), or (6.37) in terms of function values, we obtain a set of formulae expressing the integral as a sum of multiples of the values of the integrand and, in the case of formula (6.21), its derivatives. Such a formula is sometimes called a 'Lagrange-type' integration formula, by analogy with the form of Lagrange's interpolation formula.

For example, if in (6.20) we substitute

$$\mu\delta f_0 = \frac{1}{2}(f_1 - f_{-1}), \quad \mu\delta f_n = \frac{1}{2}(f_{n+1} - f_{n-1})$$

and neglect terms beyond these, we obtain

$$\int_{x_0}^{x_n} f(x) dx = (\delta x) \left[-\frac{1}{24}f_{-1} + \frac{1}{2}f_0 + \frac{25}{24}f_1 + f_2 + \dots + f_{n-2} + \frac{25}{24}f_{n-1} + \frac{1}{2}f_n - \frac{1}{24}f_{n+1} \right]. \quad (6.38)$$

Coefficients in a number of formulae of this type, differing in the order of differences to which they are correct and in the number of function values outside the range x_0 to x_n used, are given in *Chambers's 6-Figure Tables*.‡

† For an example of the great advantage of a central-difference formula over Gregory's formula, see *Chambers's 6-Figure Tables* (1949), vol. 2, pp. xxxiv and 548, or *Chambers's Shorter 6-Figure Tables* (1950), p. xxv. (The central-difference formula is there called Gauss's formula.)

‡ Vol. 2, p. 549.

The writer's own preference is for formulae in terms of differences, such as (6.17), as it is much easier to see which differences have to be taken into account, and inclusion of an extra one does not mean altering the whole formula.

6.53. Use of Simpson's or Weddle's rules

The $\delta^2 f$ terms in the integration formula (6.23) can be taken into account by using Simpson's rule for the intervals δx taken in pairs, instead of by using the term $(\mu\delta f_n - \mu\delta f_0)$ in (6.20). There is no great advantage in this procedure, except the smaller coefficient of the $\delta^4 f$ term, and it has certain mild disadvantages. It is equivalent to calculating a correction to the trapezoidal formula for *each* pair of intervals in the integration, which is unnecessary when only a single value of an integral between fixed limits is required, and it involves a substantial amount of work in calculating a correction which may vanish identically, as, for example, in $\int_0^\infty e^{-x^2} dx$. It also gives different weights to alternate function values, and requires that the total number of intervals required to cover the range of x should be even. Use of Weddle's rule has similar disadvantages, of which the fact that the number of intervals has to be a multiple of 6 may be more serious.

6.54. Integrals of functions for which $f^{(2n+1)}(x) = 0$ at both ends of the range of integration

If all odd derivatives of $f(x)$ are zero at one of the limits of $\int_a^b f(x) dx$, then the total contribution from that end of the range of integration to the correction to the trapezium rule formula, as expressed by the series of which the first few terms are given by the Euler-Maclaurin formula (6.22), is zero. And if all the odd derivatives are zero at both ends of the range, it would appear from this that the expression for the integral as the sum of a number of trapezoidal contributions is exact, whatever interval length (δx) is used in the integration.

An example is provided by the integral $(1/\pi) \int_0^\pi \cos(y \sin \theta) d\theta$ for the Bessel function $J_0(y)$. Here the integrand is an even function both of θ and of $(\pi - \theta)$, so $f^{(2n+1)}(0) = f^{(2n+1)}(\pi) = 0$ for all n , and for integration intervals $\delta\theta = \pi/N$ (N integral), the Euler-Maclaurin formula might appear to give

$$(1/\pi) \int_0^\pi \cos(y \sin \theta) d\theta = (1/2N) \left[1 + 2 \sum_{m=1}^{N-1} \cos\{y \sin(m\pi/N)\} \right] \quad (6.39)$$

exactly, for any value of N . Evaluation of the right-hand side of (6.39) for $y = 2$ and different values of N gives the following results, to the accuracy of eight-decimal tables:

| N | $\delta\theta$ | $(1/2N) \left[1 + 2 \sum_{m=1}^{N-1} \cos\{2 \sin(m\pi/N)\} \right]$ |
|-----|----------------|---|
| 12 | $\pi/12$ | 0.22389 078 |
| 8 | $\pi/8$ | 078 |
| 6 | $\pi/6$ | 079 |
| 4 | $\pi/4$ | 0.22393 5 |
| 3 | $\pi/3$ | 0.22630 |
| 2 | $\pi/2$ | 0.29193 |

The value of the integral is $J_0(2) = 0.223890779$. Thus the relation (6.39) is nearly satisfied for values of $\delta\theta$ which are quite considerable, but it is not exact; moreover the error increases with $\delta\theta$ with drastic rapidity when $\delta\theta$ is greater than about 0.6 radians.

One reason why formula (6.39) is not exact is this. Just as Taylor's series to m terms has a remainder after m terms, which remainder may not tend to zero as $m \rightarrow \infty$, so has the Euler–Maclaurin formula (6.22). If the odd derivatives of the integrand vanish at both ends of the range of integration, the remainder after m terms is independent of m ; but it does not follow that it is zero. A similar situation may occur with Taylor's series. For example, it can be shown (most easily by induction) that for the function $g(x)$ defined by

$$\begin{aligned} g(x) &= 0, & x &= 0, \\ &= e^{-1/x^2}, & x &\neq 0, \end{aligned}$$

the m th derivative at the origin, $g^{(m)}(0)$, exists and has the value zero for all values of m ; hence for every finite value of x , every term in the Maclaurin series for this function is zero, so the series converges to the value zero. Nevertheless, $g(x)$ is not zero for any value of x except $x = 0$ (though it is very small for small values of x). Similarly, the error of a trapezium-rule formula, for a given value of δx , may be non-zero although all the terms involving $(f_n^{(2m+1)} - f_0^{(2m+1)})$ in the formula for the error may be zero. This error will tend to zero as $\delta x \rightarrow 0$, but this is not the point here; the point is how the error, for a *given* value of δx , depends on the number m of 'correction' terms taken in the Euler–Maclaurin formula. A similar example is provided by the integrals $\int_0^{2\pi} f(\theta) \cos 2\pi n\theta \, d\theta$ occurring in the harmonic analysis of a periodic function (see § 11.2).

Another example is provided by the integral $\int_0^\infty e^{-x^2} \, dx$. In this case the integrand is an even function of x , so every odd derivative is zero at $x = 0$; also all derivatives are zero at $x = \infty$. Hence the Euler–Maclaurin formula appears to give

$$\int_0^\infty e^{-x^2} \, dx = (\delta x) \left[\frac{1}{2} + \sum_{j=1}^\infty e^{-j^2(\delta x)^2} \right] \quad (6.40)$$

exactly, for any value of δx . If both sides of (6.40) are evaluated for different values of δx , the results are as follows:

| | |
|--------------------------------|--|
| $\int_0^\infty e^{-x^2} \, dx$ | $= \frac{1}{2}\pi^{\frac{1}{2}} = 0.88622 \, 69254 \, 5$ to eleven decimals |
| δx | $(\delta x) \left[\frac{1}{2} + \sum_{j=1}^\infty e^{-j^2(\delta x)^2} \right]$ |
| 0.5 | 0.88622 69254 5 to eleven decimals |
| 0.6 | 69254 8 |
| 0.7 | 69285 |
| 0.8 | 0.88622 72808 |
| 0.9 | 23 598 |
| 1.0 | 32 0 |
| 1.1 | 0.88674 |

Thus the relation (6.40) is nearly true for values of δx which are quite considerable, but, like (6.39), it is not exact, and for the same reason.

Another aspect of this behaviour of the error of the trapezium rule for $\int_a^b f(x) dx$ when $f^{(2m+1)}(b) - f^{(2m+1)}(a) = 0$ is that the Euler–Maclaurin formula is only asymptotic. We saw in § 4.6 that the finite difference operators, and Taylor’s series in the form $E = e^U$, could be used freely on functions which were the products of polynomials and exponentials of *linear* functions of x ; e^{-x^2} is not of this form, and an examination of the error term is necessary before the Euler–Maclaurin formula is applied to it with δx so large that the approximation to the integrand in each interval by a sum of products of polynomials and exponentials becomes dubious. Such an examination has been carried out by Goodwin.†

The point might seem of formal rather than practical interest, since anyone with experience of numerical work, faced with the values of e^{-x^2} at intervals of, say, $\delta x = 1$, namely:

| x | $f = e^{-x^2}$ | δf | $\delta^2 f$ | $\delta^3 f$ | $\delta^4 f$ |
|-----|----------------|------------|--------------|--------------|--------------|
| 0 | 1 | | -12642 | | +30934 |
| 1 | 0.3679 | -6321 | +2825 | +15467 | -14978 |
| 2 | 0.0183 | -3496 | 3314 | +489 | -3622 |
| 3 | 0.0001 | -182 | 181 | -3133 | +2953 |
| 4 | 0.0000 | -1 | 1 | -180 | 179 |
| 5 | 0.0000 | | | 1 | 1 |

would say that these function values alone did not define the integrand well enough to justify evaluating the integral to more than two significant figures at most; it is hardly necessary actually to form the difference table to reach this conclusion.

However, the fact that the integral $\int_0^\infty e^{-x^2} dx$ has this property raises the question whether the use of relatively large intervals δx in the evaluation of other integrals of the form $\int_0^\infty e^{-x^2} f(x) dx$ may also give results of useful accuracy. This also has been examined by Goodwin.†

Further light is thrown on this by considering the finite difference integration formulae (6.20), (6.21) in terms of the cardinal function (§ 5.91). It has been mentioned in § 6.2 (following formula (6.3)) that the coefficients in the integration formula (6.2) are the integrals of corresponding coefficients in Bessel’s interpolation formula. Let these be

$$\beta_{2n} = \int_0^1 B_{2n}(\theta) d\theta.$$

Now if $f(x)$ is finite for all x , and tends to zero, as $x \rightarrow \infty$, fast enough for the series (5.41) for the cardinal function to converge, then the function whose intermediate values, between the tabular values $f(x_j)$, are given by Bessel’s interpolation formula is the cardinal function associated with these tabular values, so this cardinal function is the function whose integral is given by the infinite series

$$\frac{1}{2}(\delta x) \left[f_0 + f_1 + \sum_{n=1}^{\infty} \beta_{2n} (\delta^{2n} f_0 + \delta^{2n} f_1) \right].$$

† E. T. Goodwin, *Proc. Camb. Phil. Soc.* **45** (1949), 241.

And it can be verified from the definition of the appropriate cardinal function that its integral is given by the expression on the right-hand side of expression (6.40).

6.55. Evaluation of a definite integral when the integrand has a singularity

In evaluating an integral $\int f(x) dx$ of which the integrand has a singularity, the singularity can often be removed by a change of independent variable. If the integral is required as a function of the upper limit, we may want to avoid this in order to obtain directly, without further interpolation, values of the integral at equally spaced values of x . But this does not apply to an integral between fixed limits, and in this context the only reason for avoiding a change of variable is that a certain amount of interpolation may be required in order to obtain the values of the integrand at equal spacings in the new variable.

There is, of course, no need to use the new independent variable over the whole range. For example, to evaluate $\int_0^1 [f(x)/(1-x)^{\frac{1}{2}}] dx$ we might use the substitution $x = 1-y^2$ over the whole range of x , and so evaluate the integral as

$$\int_0^1 \frac{f(x)}{(1-x)^{\frac{1}{2}}} dx = 2 \int_0^1 f(1-y^2) dy,$$

or we might divide the range of x into two parts, one from $x = 0$ to ξ and the other from $x = \xi$ to 1 and only make the substitution in the second part, thus evaluating the integral in the form

$$\int_0^1 \frac{f(x)}{(1-x)^{\frac{1}{2}}} dx = \int_0^{\xi} \frac{f(x)}{(1-x)^{\frac{1}{2}}} dx + 2 \int_0^{\sqrt{1-\xi}} f(1-y^2) dy.$$

In this case ξ should be chosen so that $1-\xi$ is a perfect square. For example $\xi = 0.64$ might be taken; this would enable intervals of 0.04 or 0.08 in x to be used in the first integral and intervals of 0.1 in y in the second.

6.56. Definite integrals which are functions of a parameter

An important class of integrals between fixed limits comprises those which define a function of a parameter which occurs in the integrand, such as

$$\int_{-1}^1 \frac{e^{-xu}}{(1-u^2)^{\frac{1}{2}}} du \quad \text{or generally} \quad g(x) = \int_a^b f(x, u) du. \quad (6.41)$$

Such an integral can be evaluated by quadrature for each value of x ,

and this may be the only way of evaluating it. But another method of treatment may be much easier and less laborious if it is possible at all. This consists of finding a differential equation which the integral (6.41) satisfies, and solving this differential equation by a numerical process (see Chapter VII). The amount of work required to obtain a single value of the integral is then very much less than that required to carry out the evaluation by quadrature, and probably evaluation by quadrature will only be carried out for two or three values of x , to give initial values for the integration and to provide an overall check. It is not always possible to obtain such a differential equation, but many integrals of this kind of which the values are actually wanted in various contexts do satisfy differential equations. In such cases, the differential equations can often be obtained by one or two differentiations with respect to x and an integration by parts with respect to u .

Consider, for example, the function

$$f(x) = \int_0^{\infty} [e^{-u^2}/(u+x)] du,$$

which has been studied by Goodwin and Staton.[†] The range of integration here is infinite, but for $x > 0$ the integrand is of such a form that differentiation with respect to x is justified. One differentiation gives

$$f'(x) = - \int_0^{\infty} e^{-u^2}/(u+x)^2 du$$

and integration by parts with respect to u then gives

$$\begin{aligned} f'(x) &= \int_0^{\infty} e^{-u^2} \frac{d}{du} [1/(u+x)] du \\ &= [e^{-u^2}/(u+x)]_{u=0}^{\infty} + \int_0^{\infty} 2u[e^{-u^2}/(u+x)] du \\ &= -(1/x) + 2 \left[\int_0^{\infty} e^{-u^2} du - x \int_0^{\infty} \{e^{-u^2}/(u+x)\} du \right] \\ &= -(1/x) + \pi^{\frac{1}{2}} - 2xf(x), \end{aligned}$$

so that this function $f(x)$ satisfies the differential equation

$$f' + 2xf = \pi^{\frac{1}{2}} - (1/x). \quad (6.42)$$

Evaluation of $f(x)$ by quadrature for one value of x , say $x = 1$, could be used to give a value from which the numerical integration of equation (6.42) could be started, though Goodwin and Staton actually used a series expansion to obtain such a value.

[†] *Quart. J. Mech. and Applied Math.* **1** (1948), 319.

6.6. Use of unequal intervals of the independent variables

As already mentioned, in evaluating integrals between fixed limits there is no need to use values of the integrand at equal intervals in x , and there may be advantages in using formulae in terms of some other set of values. An integration formula using a finite number of values of the integrand can be regarded as giving a weighted mean of these values:

$$\int_a^b f(x) dx = (b-a) \left[\sum_k w_k f(x_k) \right], \quad (6.43)$$

where the w_k 's are the weights assigned to the values of the integrand at the points x_k . Given any $(n+1)$ points x_k , not necessarily equally spaced, values of w_k can be obtained which will make such a formula correct for any polynomial of degree up to n . And it is possible to put a condition on the w_k 's (such as that they should all be equal) and determine the corresponding x_k 's such that this formula should be exact for polynomials of degree up to n . But if no condition is imposed on either the x_k 's or the w_k 's, then these can be determined so that formula (6.43) with $(n+1)$ terms in the sum will be exact for any polynomial of degree $(2n+1)$. Such a formula is known as an $(n+1)$ -point Gaussian integration formula.

6.61. Gaussian integration formulae

The values x_k of the independent variable, at which the integrand values in a Gaussian formula are to be taken, and the weights w_k to be assigned to them in formula (6.43), can be found as follows.

By the transformation

$$\xi = [2x - (a+b)] / (b-a)$$

the range of integration is reduced to the range $\xi = -1$ to $+1$. Let $P_n(\xi)$ be the Legendre polynomial of degree n ; these polynomials have the property

$$\int_{-1}^1 P_m(\xi) P_n(\xi) d\xi = 0 \quad \text{if } m \neq n,$$

which is expressed by calling two such polynomials, of different degree, 'orthogonal' over the range $\xi = -1$ to $+1$.

Any polynomial of degree $(2n+1)$, say $p_{2n+1}(\xi)$, can be expressed as

$$p_{2n+1}(\xi) = P_{n+1}(\xi) q_n(\xi) + r_n(\xi), \quad (6.44)$$

$q_n(\xi)$ and $r_n(\xi)$ being the quotient and remainder polynomials on division

of $p_{2n+1}(\xi)$ by $P_{n+1}(\xi)$; $q_n(\xi)$ is a polynomial of degree n and r_n of degree n at most. Then

$$\int_{-1}^1 p_{2n+1}(\xi) d\xi = \int_{-1}^1 P_{n+1}(\xi) q_n(\xi) d\xi + \int_{-1}^1 r_n(\xi) d\xi. \quad (6.45)$$

Now since $q_n(\xi)$ is a polynomial of degree n , it can be expressed as a linear combination of Legendre polynomials $P_m(\xi)$ with $m \leq n$; each of these is orthogonal to $P_{n+1}(\xi)$, hence

$$\int_{-1}^1 P_{n+1}(\xi) q_n(\xi) d\xi = 0, \quad (6.46)$$

whatever the quotient polynomial $q_n(\xi)$, and (6.45) becomes

$$\int_{-1}^1 p_{2n+1}(\xi) d\xi = \int_{-1}^1 r_n(\xi) d\xi; \quad (6.47)$$

the vanishing of the integral (6.46) through the orthogonal property of the Legendre polynomials is the reason for taking the Legendre polynomial $P_{n+1}(\xi)$ as the divisor in (6.44).

Now we want to find values of ξ_k such that

$$\int_{-1}^1 p_{2n+1}(\xi) d\xi = 2 \sum_k w_k p_{2n+1}(\xi_k),$$

that is, on substitution from (6.44),

$$\int_{-1}^1 p_{2n+1}(\xi) d\xi = 2 \sum_k w_k [P_{n+1}(\xi_k) q_n(\xi_k) + r_n(\xi_k)]. \quad (6.48)$$

The result (6.47) shows that the integral on the left-hand side of (6.48) is independent of the quotient polynomial $q_n(\xi)$, and the expression on the right-hand side of (6.48) can be made independent of $q_n(\xi)$ if (and only if) the ξ_k 's are taken as the roots of the polynomial equation

$$P_{n+1}(\xi) = 0. \quad (6.49)$$

These roots are all real and distinct, and lie in the range $-1 < \xi < 1$; they give the values of ξ at which the values of the integrand are to be taken.

Since $r_n(\xi)$ is a polynomial of degree n , it is determined completely by its values for $(n+1)$ distinct values of ξ . Let these be taken as the $(n+1)$ roots ξ_k ($1 \leq k \leq n+1$) of equation (6.49), and let $F_{n+1}(\xi)$ be the function

$$F_{n+1}(\xi) = (\xi - \xi_1)(\xi - \xi_2) \dots (\xi - \xi_{n+1}). \quad (6.50)$$

Then from the expansion of $r_n(\xi)/F_{n+1}(\xi)$ in partial fractions (or from the equivalent Lagrange interpolation formula, see § 5.7),

$$r_n(\xi) = \sum_k \frac{F_{n+1}(\xi)}{F'_{n+1}(\xi_k)(\xi - \xi_k)} r_n(\xi_k). \quad (6.51)$$

This formula is exact, since $r_n(\xi)$ is a polynomial of degree n , and integration gives

$$\int_{-1}^1 r_n(\xi) d\xi = \sum_k \int_{-1}^1 \frac{F_{n+1}(\xi) d\xi}{F'_{n+1}(\xi_k)(\xi - \xi_k)} r_n(\xi_k). \quad (6.52)$$

Now from formulae (6.47), (6.48), the weights w_k are given by

$$\int_{-1}^1 r_n(\xi) d\xi = 2 \sum_k w_k r_n(\xi_k),$$

and comparison of this with (6.52) gives for the weights w_k the values

$$w_k = \frac{1}{2} \int_{-1}^1 \frac{F_{n+1}(\xi) d\xi}{F'_{n+1}(\xi_k)(\xi - \xi_k)}.$$

Values of w_k and ξ_k for Gaussian quadrature formulae up to $n = 16$ have been tabulated by Lowan, Davids, and Levenson.† The limits of integration can be reduced to 0 and 1, instead of -1 and 1, by the transformation $\eta = (x-a)/(b-a) = \frac{1}{2}(1+\xi)$; values of η_k and w_k are given by Whittaker and Robinson.‡

Example: To evaluate $\int_0^{0.8} e^{+x^2} dx$ by a five-point Gauss formula. The values of ξ_j , w_j , x_j , and the integrand values are as follows:

| j | ξ_j | w_j | x_j | $f_j = \exp x_j^2$ |
|---------------------------------|----------|----------|---------------|--------------------|
| 1 | 0.046910 | 0.118464 | 0.037528 | 1.001409 |
| 2 | .230765 | .239314 | .184612 | 1.034669 |
| 3 | .5 | .284444 | .400000 | 1.173511 |
| 4 | .769235 | .239314 | .615388 | 1.460388 |
| 5 | 0.953090 | 0.118464 | 0.762472 | 1.788475 |
| $\sum w_j f_j = 1.261401$ | | | $(b-a) = 0.8$ | |
| $(b-a) \sum w_j f_j = 1.009121$ | | | | |

Note: This value of the integral agrees with those calculated in §§ 6.4 and 6.5.

For work on numerical evaluation of integrals of given functions, the practical value of a Gaussian formula such as (6.43) is limited by two

† *Bull. Amer. Math. Soc.* **48** (1942), 739. See also Kopal, *Numerical Analysis* (Chapman and Hall, 1955), appendix iv, § 4.1.

‡ *Calculus of Observations* (Blackie, 4th ed., 1944), § 80.

things: first, the interpolation which may be necessary to obtain the values of the integrand at the required values of x_k , and secondly the difficulty of checking adequately both these values, which are at unequal intervals in x , and the evaluation of the integral from them. Further, if the integrand is known at equal intervals of x , as will often happen, no advantage is taken of this. When the integrand is given by a formula sufficiently simple for its value to be calculated from this formula for each value of x_k , the need for interpolation does not arise, and in such a context use of a Gaussian formula may be a practicable and useful process. A check can be provided by carrying out two independent integrations, say one with an n -point formula and the other with an $(n+2)$ -point formula.

Use of a Gaussian integration formula may also be very valuable in simplifying problems in more than one variable. It can be used, for example, to simplify integro-differential equations involving integrals

of the type $\int_{\theta=0}^{\pi} f(r, \theta) \sin \theta \, d\theta$. If such an integral is replaced by a sum $\sum_j w(\theta_j) f(r, \theta_j)$, the solution of the equation in which the integral occurs can be reduced to the solution of a finite number of equations for the functions $f(r, \theta_j)$, each of which is a function of the single variable r only. In making such a replacement it is clearly desirable to obtain as good an approximation as possible with a small number of terms, and this is obtained by taking the values of θ_j and the weights $w(\theta_j)$ to be those of a Gauss formula.†

6.62. Gaussian formula for $\int_0^{\infty} e^{-kx} p_{2n+1}(x) \, dx$

The argument of the previous section will not apply to an integral over an infinite range. However, a similar argument can be applied to integrals of the type $\int_0^{\infty} e^{-kx} p_{2n+1}(x) \, dx$ and $\int_0^{\infty} e^{-kx^2} p_{2n+1}(x) \, dx$, where $p_{2n+1}(x)$ is a polynomial of degree $(2n+1)$ in x , and the results can then be applied to the approximate evaluation of infinite integrals in which the integrand, though not exactly of the form $e^{-kx} p_{2n+1}(x)$ or $e^{-kx^2} p_{2n+1}(x)$, is known or believed to be approximately of one of these forms. The coefficient k in the exponential factor can be removed by a change of scale of the independent variables; we shall consider as an example the integral $\int_0^{\infty} e^{-x} p_{2n+1}(x) \, dx$.

The argument leading to formulae (6.49), (6.52) of the previous section depended

† See, for example, S. Chandrasekhar, *Astrophys. Journ.* **100** (1944), 76, and *Radiative Transfer* (Clarendon Press, 1950), §§ 20, 25; also G. C. Wick, *Zeit. f. Phys.* **121** (1943), 702.

on the use of the orthogonal property, of the Legendre polynomials to give the result (6.46). Correspondingly we now want to write

$$p_{2n+1}(x) = L_{n+1}(x)q_n(x) + r_n(x),$$

where the divisor polynomial $L_{n+1}(x)$ is of degree $(n+1)$ and is one of a set of polynomials such that

$$\int_0^\infty e^{-x} L_m(x) L_n(x) dx = 0 \quad \text{for } m \neq n, \quad (6.53)$$

that is, the functions $e^{-x} L_n(x)$ are orthogonal over the range $x = 0$ to ∞ . Then, corresponding to (6.46) we will have $\int_0^\infty e^{-x} L_{n+1}(x) q_n(x) dx = 0$, and consequently

$$\int_0^\infty e^{-x} p_{2n+1}(x) dx = \int_0^\infty e^{-x} r_n(x) dx,$$

corresponding to (6.47). Apart from the presence of the factor e^{-x} in this integrand, the rest of the argument follows that of § 6.61.

The polynomials $L_n(x)$ with the orthogonal property (6.53) are those known as the Laguerre polynomials, defined by

$$L_n(x) = e^{+x} \left(\frac{d}{dx} \right)^n (x^n e^{-x});$$

for an n -point formula (correct for polynomials $p_{2n+1}(x)$), the values of the x_k 's are the roots of

$$L_{n+1}(x) = 0,$$

and if

$$F_{n+1}(x) = (x-x_1)(x-x_2)\dots(x-x_{n+1}),$$

the weights are

$$w_k = \int_0^\infty \frac{F_{n+1}(x) e^{-x} dx}{F'_{n+1}(x_k)(x-x_k)}.$$

Values of x_k, w_k for n -point formula up to $n = 15$ have been calculated by Salzer and Zucker.†

In applying these results to the approximate evaluation of integrals $\int_0^\infty f(x) dx$ in which the formal behaviour of the integrand is not precisely known, it is necessary to take a factor e^{-kx} out of the integrand $f(x)$, and the result may depend on the value of k adopted. For this reason, some discretion is required in such a context.

6.7. Numerical differentiation

We have already seen that a table of values does not define a function uniquely. Still less does it establish whether the function tabulated is differentiable everywhere, or even anywhere, within the range of the table; two functions may be indistinguishable, to any specified degree of numerical accuracy, for every value of x (not only for the tabular

† *Bull. Amer. Math. Soc.* **55** (1949), 1004. See also Z. Kopal, *Numerical Analysis* (Chapman and Hall, 1955), appendix iv, § 4.2. Tables of abscissae and weights for a number of other integration formulae of Gauss type are also given by Kopal.

values), yet one may be differentiable everywhere and the other nowhere. And still less does a table establish whether a function is differentiable two or more times. These considerations alone suggest that numerical differentiation of a function specified by a table may be a dubious process.

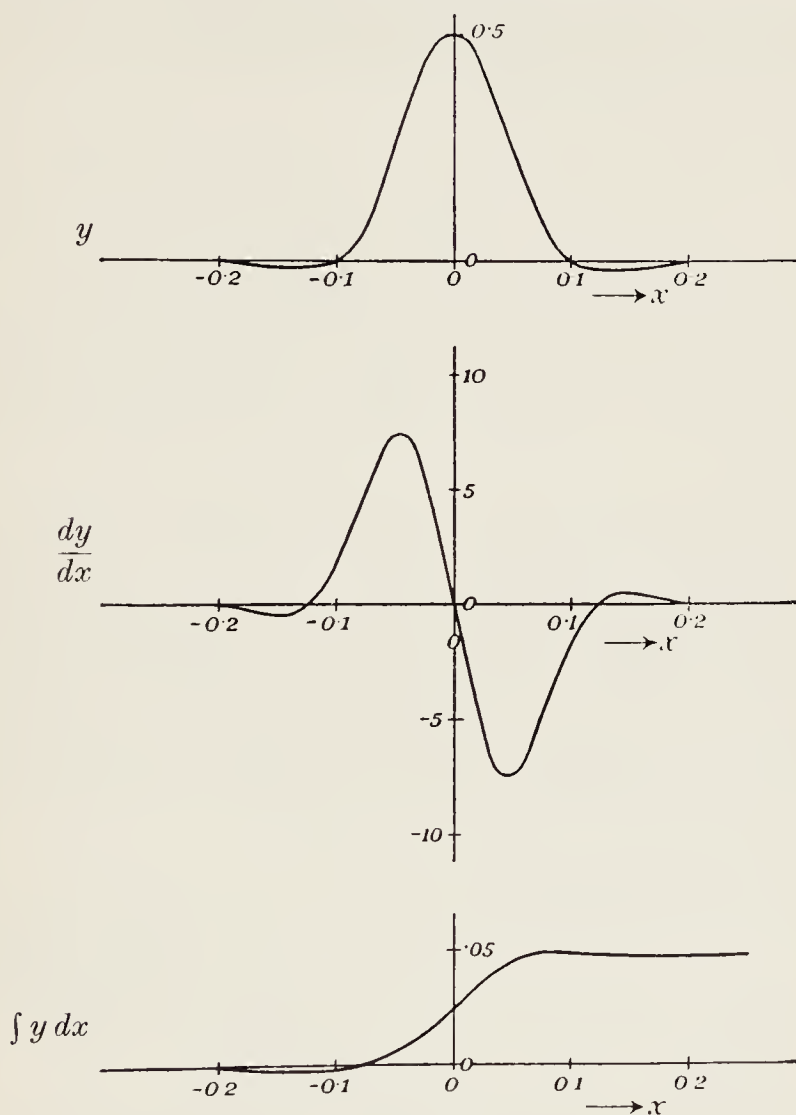


FIG. 10.

Further, the entries in a table are affected in an irregular way by rounding errors, and differentiation exaggerates irregularities whereas integration smooths them. In differentiation we are concerned with a limit process carried out on the quantity $[f(x+\delta x)-f(x)]/(\delta x)$, and as δx becomes smaller, irregularities in the values of f become *more* prominent in the result; whereas in integration we are concerned with a limit process carried out on the quantity $\sum f \delta x$ and the effect of an

irregularity in any one value of f becomes *less* prominent as δx becomes smaller. This is illustrated by Fig. 10, which shows the graph of the function

$$y = \frac{1}{2} \frac{\sin 10\pi x}{10\pi x} e^{-100x^2},$$

which might represent approximately an irregularity produced by a rounding error at $x = 0$ in a table at intervals of 0.1 in x , and the derivative and integral of this function.

For such reasons, the differentiation of a function specified only by a table of values, or determined experimentally and so subject to errors of observation, is a notoriously unsatisfactory process, particularly if higher derivatives than the first are required. It is a process to be avoided if possible, unless the context in which the results are required is such that the limited accuracy attainable by the numerical process is certainly adequate. In general, values of the second and higher derivatives obtained from such data should be regarded with caution if not scepticism.

In some cases it may be possible to evaluate derivatives by some process other than numerical differentiation. If, for example, a function y is known to satisfy a first-order differential equation, values of y' can be obtained by substituting the values of y into the differential equation. And if y satisfies a differential equation of higher order, it is usually better to obtain y' by numerical solution of the equation as an equation for y' , taking y as a given function of x , than to try to obtain y' directly from the values of y by a differentiation formula.

6.71. Differentiation formulae

To obtain a first-order derivative there are two useful formulae, one giving the values of the derivative f' at the values x_j of x at which the function is tabulated, and the other giving f' at $x_{j+\frac{1}{2}}$. The former has already been obtained in § 4.71, and is (see formula (4.46))

$$\begin{aligned} (\delta x)f'_0 &= \mu\delta f_0 - \frac{1}{6}\mu\delta^3 f_0 + \frac{1}{360}\mu\delta^5 f_0 - \frac{1}{1440}\mu\delta^7 f_0 + O(\delta x)^9 \\ &= \frac{1}{2}[(f_1 - f_{-1}) - \frac{1}{6}(\delta^2 f_1 - \delta^2 f_{-1}) + \frac{1}{360}(\delta^4 f_1 - \delta^4 f_{-1}) - \frac{1}{1440}(\delta^6 f_1 - \delta^6 f_{-1})] + O(\delta x)^9. \end{aligned} \quad (6.54)$$

The other can be obtained either by differentiating Bessel's interpolation formula with respect to p and then putting $p = \frac{1}{2}$, or by using finite-difference operators as follows. We want to find a $\phi(\delta)$ such that

$$(\delta x)f'_{\frac{1}{2}} = \phi(\delta)\delta f_{\frac{1}{2}}.$$

Hence

$$\phi(\delta) = U/\delta = (\sinh^{-1} \frac{1}{2}\delta)/\frac{1}{2}\delta = 1 - \frac{1}{24}\delta^2 + \frac{3}{640}\delta^4 - \frac{5}{7168}\delta^6 + O(\delta x)^8$$

on putting $n = 1$ in (4.42). Hence

$$(\delta x)f'_\frac{1}{2} = \delta f_\frac{1}{2} - \frac{1}{24}\delta^3 f_\frac{1}{2} + \frac{3}{640}\delta^5 f_\frac{1}{2} - \frac{5}{7168}\delta^7 f_\frac{1}{2} + O(\delta x)^9. \quad (6.55)$$

This formula is much preferable to (6.54) on account of the more rapid decrease of the coefficients of the higher orders of differences. If, however, values of f'_j are required, there is no advantage in using (6.55) followed by 'half-way' interpolation between the values of $f'_{j+\frac{1}{2}}$ by use of formula (5.3), since these two processes together just give formula (6.54).

For a second-order derivative, the appropriate formula is (4.43), namely

$$(\delta x)^2 f''_0 = \delta^2 f_0 - \frac{1}{12}\delta^4 f_0 + \frac{1}{90}\delta^6 f_0 - \frac{1}{560}\delta^8 f_0 + O(\delta x)^{10}. \quad (6.56)$$

In carrying out the calculations, the interval (δx) taken *should not be too small*, since the smaller it is taken, the smaller the number of significant figures in $\delta f_\frac{1}{2}$ and so in f' (and similarly for a second derivative). Rather, δx should be taken as large as is convenient, subject to the truncation error of the differentiation formula used being negligible.†

It will often be advisable either to smooth the values of f before differentiation, or to smooth the values of f' or f'' obtained (for a smoothing process see § 11.4). Let us write f'_s for the smoothed values which form an approximation to f' . To ensure that no systematic errors are introduced in the smoothing process, the values of f'_s should be integrated, and compared with the original values of f . If the quantities $f - \int f'_s dx$, which are called the 'residuals', show any significant systematic variation with x , a process of differentiation should be carried out on these residuals.

Example: The function tabulated below as $Y(x)$ is the solution of the equation $y'' = 1 + xy$ with $y(0) = y'(0) = 0$, and the function $z(x)$ is the solution of $z'' = xz$ with $z(0) = 0$, $z'(0) = 1$; to find $y'(0)$ and $y(2)$ for the solution of $y'' = 1 + xy$ for which $y(0) = 0$, $y'(2) = 0$.

| x | $Y(x)$ | $\delta^2 Y$ | $\delta^4 Y$ | $\delta^6 Y$ | $z(x)$ |
|-----|---------|--------------|--------------|--------------|-----------|
| 1.6 | 1.56205 | | | | |
| 1.7 | 1.83254 | 4125 | | | |
| 1.8 | 2.14428 | 4873 | 154 | | 2.80444 |
| 1.9 | 2.50475 | 5775 | 188 | 7 | 3.17749 |
| 2.0 | 2.92297 | 6865 | 229 | 7 | 3.61107 |
| 2.1 | 3.40984 | 8184 | 277 | 18 | 4.11708 |
| 2.2 | 3.97855 | 9780 | 343 | | 4.70978 |
| 2.3 | 4.64506 | 11719 | | | |
| 2.4 | 5.42876 | | | | |
| | | | | | $z'(2.0)$ |
| | | | | | = 4.67626 |

† For a closer analysis of the best interval to use in numerical differentiation, see Z. Kopal, *Numerical Analysis* (Chapman and Hall, 1955), § III-E.

Two standard solutions of the equation $y'' = xy$ have been tabulated.† They are written $\text{Ai}(x)$ and $\text{Bi}(x)$, and the function $z(x)$ of this example is related to them by

$$z(x) = [\text{Bi}(x) - 3^{\frac{1}{2}}\text{Ai}(x)]/2 \cdot 3^{\frac{1}{2}}\beta,$$

where β is a constant given in the Introduction to the Tables (p. B. 17; the value of $2 \cdot 3^{\frac{1}{2}}\beta$ is 0.896577; this function $z(x)$ is that written $y_2(x)$ in that Introduction). The values of $z(x)$ and $z'(x)$ here tabulated have been calculated from this formula; only the value of $z(2.0)$ is required to give the results sought, but neighbouring values are given for use in a check.

The general solution of $y'' = 1 + xy$ with $y = 0$ at $x = 0$ is

$$y = Y + cz,$$

where c is an arbitrary constant; for the solution with $y'(2.0) = 0$,

$$c = -Y'(2.0)/z'(2.0),$$

so we need to determine $Y'(2.0)$. From the tabulated results and formula (6.43)

$$\begin{aligned} 0.2Y'(2.0) &= 0.90509 - \frac{1}{8}(2409) + \frac{1}{36}(89) - \frac{1}{144}(11) \\ &= 0.90509 - 401_5 + 3_0 - 0_1 = 0.90110_4, \end{aligned}$$

while $0.2z'(2.0) = 0.935252$, so $c = -(0.90110_4)/(0.935252) = -0.96348_2$.

With this value of c we have

| x | Y | $-cz$ | y | $\delta^2 y$ | $\delta^4 y$ |
|-----|---------|----------|----------|--------------|--------------|
| 1.8 | 2.14428 | -2.70203 | -0.55775 | | |
| 1.9 | 2.50475 | -3.06145 | -0.55670 | - 58 | |
| 2.0 | 2.92297 | -3.47920 | -0.55623 | -113 | - 1 |
| 2.1 | 3.40984 | -3.96673 | -0.55689 | -169 | |
| 2.2 | 3.97855 | -4.53779 | -0.55924 | | |

and for this function y , $y(2.1) - y(1.9) = -0.00019$ and $\delta^2 y(2.1) - \delta^2 y(1.9) = -111$,

$$\begin{aligned} \text{so } 0.2y'(2.0) &= -0.00019 - \frac{1}{8}(-111) + \frac{1}{36}(-1) \\ &= -0.00019 + 18_5 = -0.00000_5, \end{aligned}$$

which is within the tolerance for rounding errors. This provides a check of the work.

Notes: (i) The value of $y(2.0)$ is not determined correctly to a unit in the last figure; the value $c = -0.96347_5$ gives $0.2y'(2.0) = +0.00000_5$, which is equally within the tolerance for rounding errors, and $y(2.0) = -0.55621$.

(ii) The value of $y'(0)$ for this solution is just $y'(0) = c$.

(iii) The value of $Y'(2.0)$ is not determined with certainty to several units in the fifth decimal, since $y(2.1) - y(1.9)$ is subject to rounding errors up to 1 in the fifth decimal and is multiplied by $1/2(\delta x) = 5$. A more accurate value of $Y'(2.0)$ could be obtained by using $\delta x = 0.2$, but the contribution to $2(\delta x)Y'$ from the higher orders of differences would be considerably greater; that from $\delta^6 Y$, for example, would be greater by a factor of over 100.

† *British Association Mathematical Tables*, Part-volume B (1946), *The Airy Integral*.

6.72. Graphical differentiation

The residuals $f - \int f'_s dx$ of a numerical differentiation over a range of x will usually be numbers of one or two digits only, so can easily be plotted to the accuracy to which they are known. In such a case, a graphical method of carrying out the differentiation is adequate. The best way of doing this is to plot on *good* squared paper (see § 2.5) the values of the function to be differentiated, and through each plotted point draw a vertical line to indicate the range of uncertainty, due to rounding error or other causes, of that value. Then draw the smoothest curve passing each plotted point within the indicated tolerance. The latitude in drawing such a curve will give an indication of the reliability of the values of the derivative.

The best way of finding the gradient of such a curve, or of one representing a set of results of some experiment or observations, is as follows. Take a flat piece of polished sheet metal (aluminium or stainless steel is satisfactory), or surface-aluminized glass, mounted in such a way that it can be placed on a piece of paper with its surface accurately perpendicular to the paper and extending right down to the paper. Set this so as to intersect the curve at the point at which the gradient is wanted (see Fig. 11), and rotate it until there is no discontinuity in direction between the curve and its reflection in the mirror. With care, this setting can be made with considerable accuracy, probably greater than that to which the curve can be drawn. The gradient of the curve can then be determined directly from the intersections of the plane of the mirror with the grid lines of the paper in which the curve is plotted.

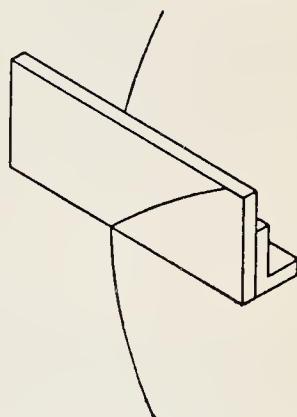


FIG. 11.

6.8. Errors of interpolation and integration formulae

W. E. Milne† has given a convenient general method for obtaining formally exact expressions for the truncation errors of formulae for interpolation, integration, etc.

We have derived and expressed such formulae as results of linear operations on the function to be interpolated, integrated, or differentiated. If we take one of these formulae to a finite number of terms, then the remainder after n terms can also be expressed as the result of a linear operation on this function. For example, if we take formula (6.11) as

† *Numerical Calculus* (Univ. of Princeton Press, 1949), §§ 30, 31.

far as the terms in f' , the remainder, which is the truncation error of the formula in this form, is

$$\begin{aligned} \int_{x_0}^{x_1} f dx - \frac{1}{2}(\delta x)[(f_0 + f_1) - \frac{1}{6}(\delta x)(f'_1 - f'_0)] \\ = (\delta x) \left[\frac{1}{U}(E-1) - \frac{1}{2}(E+1) + \frac{1}{12}(E-1)U \right] f_0, \quad (6.57) \end{aligned}$$

which is the result of a linear operation R on f . Milne calls an operator R 'of degree n ' when $Rx^m = 0$ for $m \leq n$, $Rx^{n+1} \neq 0$, and writes R_n for an operator of degree n . The purpose is to obtain an expression for $R_n f$ for any function f . It is assumed that R_n does not involve higher powers of U than U^{n-1} .

R may contain some shift operators E so that Rf_0 may depend on values of f or its derivatives for values of x other than x_0 . Let x_m be the least and x_M the greatest of these; and let a be a value of x less than x_m . Also let us write $\phi_n(z)$ for the function

$$\phi_n(z) = z^n \quad \text{for } z \geq 0, \quad \phi_n(z) = 0 \quad \text{for } z < 0. \quad (6.58)$$

One form of Taylor's series to $n+1$ terms with a remainder is

$$\begin{aligned} f(x) = f(a) + (x-a)f'(a) + \frac{1}{2!}(x-a)^2 f''(a) + \dots + \\ + \frac{1}{n!}(x-a)^n f^{(n)}(a) + \frac{1}{n!} \int_a^x f^{(n+1)}(\xi)(x-\xi)^n d\xi; \quad (6.59) \end{aligned}$$

this form can be obtained by repeated integration by parts, using

$$m \int_a^x f^{(m)}(\xi)(x-\xi)^{m-1} d\xi = (x-a)^m f^{(m)}(a) + \int_a^x f^{(m+1)}(\xi)(x-\xi)^m d\xi.$$

The last term in (6.59) can be written

$$\frac{1}{n!} \int_a^\infty f^{(n+1)}(\xi) \phi_n(x-\xi) d\xi, \quad (6.60)$$

since the integrand here is zero for $\xi > x$.

The first $n+1$ terms in (6.59) form a polynomial of degree n , so they are annihilated by the operator R_n . Also since R_n operates on functions in so far as they are functions of x , it only operates on the function ϕ_n in the integral in the form (6.60); this is the reason for expressing the

integral in this form, in which the limits are independent of x . Hence

$$R_n f(x) = \frac{1}{n!} \int_a^\infty f^{(n+1)}(\xi) R_n \phi_n(x-\xi) d\xi.$$

Since a has been chosen to be smaller than the smallest argument x_m occurring in $R_n f$, it follows that the arguments of all the terms in $R_n \phi_n(x-\xi)$ are positive for $\xi < a$; but $\phi_n(z) = z^n$ for positive z , so $R_n \phi_n(x-\xi) = 0$ for $\xi < a$. So the lower limit of the integral can be replaced by $-\infty$, and finally

$$R_n f(x) = \int_{-\infty}^\infty f^{(n+1)}(\xi) G(\xi) d\xi, \quad (6.61)$$

where
$$G(\xi) = \frac{1}{n!} R_n \phi_n(x-\xi). \quad (6.62)$$

If $R_n f(x)$ is a function of x , then $G(\xi)$ is a function of x as well as of ξ ; if $R_n f(x)$ is not a function of x , as is the case for the operator in (6.57), then $G(\xi)$ is not a function of x .

The function $G(\xi)$ consists of polynomial segments between the values of x_j involved in $R_n f(x)$, and is zero outside the range of these values. It is also independent of the function on which R_n operates; hence

$$|R_n f(x)| \leq K \cdot [\max |f^{(n+1)}(x)| \text{ in } x_m \leq x \leq x_M],$$

where
$$K = \int_{-\infty}^\infty |G(\xi)| d\xi;$$

K is independent of the function f on which R operates.

In many cases $G(\xi)$ is of constant sign over the range where it is not zero, and then a better formula for the error can be obtained. The mean value theorem, applied to (6.61), then gives

$$R_n f(x) = f^{(n+1)}(X) \int_{-\infty}^\infty G(\xi) d\xi, \quad (6.63)$$

where $x_m \leq X \leq x_M$; the integrand of (6.61) is zero outside these limits so X must lie in this range. Also for $f(x) = x^{n+1}/(n+1)!$, $f^{(n+1)}(x) = 1$ everywhere, so that in this case (6.63) gives

$$\int_{-\infty}^\infty G(\xi) d\xi = R_n x^{n+1}/(n+1)!$$

and hence, in general,

$$R_n f(x) = f^{(n+1)}(X) R_n x^{n+1}/(n+1)! \quad (6.64)$$

As Milne points out, the evaluation of $R_n x^{n+1}$ on the right-hand side of (6.64) is often much easier than the determination of the polynomial segments of $G(\xi)$ and their integration. In (6.64), $R_n x^{n+1}$ can be replaced by $R_n(x-b)^{n+1}$ for any constant b if this is more convenient for the evaluation of this quantity.

Example: To obtain a formula for the error of trapezium rule integration

$$\int_{x_0}^{x_1} f dx = \frac{1}{2}(\delta x)[f_0 + f_1].$$

Here
$$Rf(x) = \int_{x_0}^{x_1} f dx - \frac{1}{2}(\delta x)(f_0 + f_1),$$

which is identically zero for $f(x) = x$, but not for $f(x) = x^2$. Hence R is an R_1 , and

$$R_1 \phi_1(x-\xi) = \int_{x_0}^{x_1} \phi_1(x-\xi) dx - \frac{1}{2}(\delta x)[\phi_1(x_0-\xi) + \phi_1(x_1-\xi)].$$

For $\xi < x_0$, $x-\xi$ is positive over the whole range $x = x_0$ to x_1 , so $\phi_1(x-\xi) = x-\xi$ for all relevant ξ , and

$$1! G_1(\xi) = R_1 \phi_1(x-\xi) = \frac{1}{2}[(x-\xi)^2]_{x=x_0}^{x_1} - \frac{1}{2}(\delta x)[x_0-\xi + x_1-\xi],$$

which is zero, as it should be. For $x_0 < \xi \leq x_1$, $\phi_1(x-\xi) = 0$ for $x < \xi$, so

$$1! G_1(\xi) = R_1 \phi_1(x-\xi) = \frac{1}{2}[(x-\xi)^2]_{x=\xi}^{x_1} - \frac{1}{2}(\delta x)(x_1-\xi) = \frac{1}{2}(x_1-\xi)(x_0-\xi).$$

For $\xi > x_1$, $\phi_1(x-\xi)$ is zero over the whole range of x ; so $G_1(\xi) = 0$. Hence altogether

$$\begin{aligned} G_1(\xi) &= \frac{1}{2}(x_1-\xi)(x_0-\xi) \quad \text{for } x_0 \leq \xi \leq x_1 \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

Hence for a general function $f(x)$,

$$R_1 f = \frac{1}{2} \int_{x_0}^{x_1} f''(\xi)(x_1-\xi)(x_0-\xi) d\xi.$$

In this case $G(\xi)$ is zero or negative, and is of so simple a form that it is easy to evaluate $\int_{-\infty}^{\infty} G(\xi) d\xi$ directly. Substitution of $\xi = x_0 + (x_1 - x_0)\eta$ gives

$$\int_{-\infty}^{\infty} G(\xi) d\xi = \frac{1}{2} \int_{x_0}^{x_1} (x_1-\xi)(x_0-\xi) d\xi = -\frac{1}{2}(\delta x)^3 \int_0^1 \eta(1-\eta) d\eta = -\frac{1}{12}(\delta x)^3.$$

Alternatively, taking $f(x) = (x-x_0)^2/2!$ and using (6.64) we have

$$\begin{aligned} R_1(x-x_0)^2/2! &= \frac{1}{2}[\frac{1}{3}(x_1-x_0)^3 - \frac{1}{2}(\delta x)(x_1^2+x_0^2)] \\ &= \frac{1}{12}(\delta x)[2(x_1^2+x_1x_0+x_0^2) - 3(x_1^2+x_0^2)] \\ &= -\frac{1}{12}(\delta x)^3, \end{aligned}$$

so that $R_1 f = -\frac{1}{12} f''(X)(\delta x)^3$, where $x_0 \leq X \leq x_1$.

6.81. Use of formulae for the error

If a formula for interpolation, integration, etc., is such that the operator R of the previous section is of degree n , the error of the formula

involves the $(n+1)$ th derivative of the function to which it is applied. But we have seen that for a function specified only by a table of values, the numerical determination of derivatives beyond the first or second is an unreliable process and one to be avoided if possible. Even when the function to which the interpolation or integration formula is to be applied is given by a formula which can be differentiated, the formulae for the higher derivatives may be too complicated to be convenient for numerical evaluation. Thus a formula which depends on the values of derivatives beyond the first or second is of limited practical use.

VII

INTEGRATION OF ORDINARY DIFFERENTIAL EQUATIONS

7.1. Step-by-step methods

ONE class of methods for the numerical integration of ordinary differential equations consists of those in which the solution is evaluated step by step through a series of equal intervals in the independent variable, so that when the solution has been carried to $x = x_j$, the next step consists of evaluating the change in the solution through the interval δx from x_j to x_{j+1} . In such a process we follow out in the course of the numerical work the development of the solution as the independent variable increases. For simple equations this can be made a straightforward and easy process to carry out; it can be provided with adequate current checks to assure the worker that the integration is proceeding correctly, and in the writer's experience it is one of the most satisfying forms of numerical work to carry out.

7.11. One-point and two-point boundary conditions

From the point of view of a step-by-step process, the nature of the boundary and other conditions to be satisfied is more important than the nature of the equation itself; and as regards the boundary conditions what matters is not *what* they are but *where* they are. If all the conditions which the solution must satisfy are boundary conditions given at one point of the range of integration (usually one end of it), the solution can be started from there with all relevant quantities known; and, apart from the possible occurrence of singularities or of instability in the process of integration, evaluation of a solution usually gives no difficulty. Such conditions are known as 'one-point' boundary conditions, and a problem in which the conditions are of this type has been called by Richardson† a 'marching problem' as the solution is obtained by marching step by step from the initial data.

But if some conditions are specified at one point, $x = a$, of the range and others at another, $x = b$ (usually $x = a, b$ will be the ends of the range) or if there is a relation between the behaviour of the solution at the two ends of the range such as a condition that the solution y should be periodic, which for a first-order equation is $y(b) = y(a)$, or some

† L. F. Richardson, *Phil. Trans. Roy. Soc.* 226 (1927), 300.

integral condition on the solution as a whole such as $\int_a^b y^2 dx = 1$, the evaluation of the solution may not be so straightforward. Conditions which are specified at two points of the range are called 'two-point' boundary conditions; a set of conditions at more than two points is possible but unusual. A problem in which the conditions on the solution are not one-point boundary conditions Richardson has called a 'jury problem'.

If the equation, and the conditions which the solution must satisfy, are linear, it may be possible to evaluate the solution as the sum of a particular integral satisfying the conditions at one point of the range, and a complementary function. But in many cases such a procedure is a formal possibility only and not a useful one for practical numerical work (see § 7.6); and if the equation is non-linear it is not available.

A step-by-step solution has to start from some point of the range with definite numerical values of sufficient quantities to define a solution; for an n th-order equation these will usually be y and its first $(n-1)$ derivatives, but they may be the values $y_0, y_1, y_2, \dots, y_{n-1}$ at the beginnings of the first n intervals. With one-point boundary conditions this point is naturally taken as the point from which to start the integration. With other conditions it is best to start from the point at which the values of the greatest number of values of y or its derivatives are specified by the given conditions on the solution. The other starting conditions have to be estimated and adjusted, either by trial or by the use of a complementary function when this is practicable, until a solution satisfying the other conditions is obtained.

We will consider first the step-by-step evaluation of a solution from *given* initial conditions, and later (§ 7.6) return to the consideration of the determination of solutions satisfying other conditions.

7.2. Second-order equation with first derivative absent

The simplest numerical process is that for a second-order equation with the first derivative absent

$$y'' = f(x, y), \quad (7.1)$$

in which $f(x, y)$ need *not* be linear in y . This is integrated by using the formula for twofold integration from y'' to y , without an intermediate calculation of y' (§ 6.44):

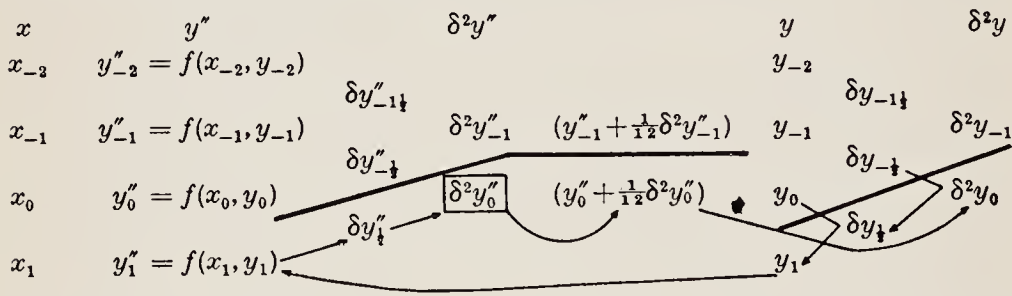
$$y_1 - 2y_0 + y_{-1} = \delta^2 y_0 = (\delta x)^2 [y_0'' + \frac{1}{12} \delta^2 y_0'' - \frac{1}{240} \delta^4 y_0''] + O(\delta x)^8. \quad (7.2)$$

One procedure will be explained first in some detail; there are several

variants of it, some of which will be mentioned later. It will be supposed that the term $\frac{1}{240}\delta^4y''_0$ in (7.2) is negligible, so that the integration formula is being used in the form

$$\delta^2y_0 = (\delta x)^2[y''_0 + \tfrac{1}{12}\delta^2y''_0]. \tag{7.3}$$

Suppose the integration has reached $x = x_0$, and we are concerned with the integration through the interval δx to $x = x_1$. At this stage we have y_0 and $y''_0 = f(x_0, y_0)$ and the *backward* differences from these. The procedure is then as follows. Estimate $\delta^2y''_0$, and obtain an approximation to δ^2y_0 from (7.3). Add this to $\delta y_{-\frac{1}{2}}$ to give an approximation to $\delta y_{\frac{1}{2}}$, and add this to y_0 to give an approximation to y_1 . From this calculate y''_1 and hence $\delta^2y''_0 = y''_1 - 2y''_0 + y''_{-1}$. Let ϵ be the difference between this value of $\delta^2y''_0$ and that estimated. A change of the estimate of $\delta^2y''_0$ by ϵ makes a change $\frac{1}{12}(\delta x)^2\epsilon$ in y_1 . If this is less than $\frac{1}{2}$ in the last figure retained in y , the estimate is adequate; if not, the estimate is revised and the calculation of the interval repeated; but the interval length (δx) should be taken so that this is seldom necessary. A convenient arrangement of the work is as follows:



The quantities above the heavy lines are those which are known when the integration has reached $x = x_0$; the quantity $\delta^2y''_0$ enclosed in a 'box' is that which is estimated and if necessary adjusted, and the arrows show the sequence in which the various quantities are calculated.

To start the integration, two values of y are required, and it is advisable to have three to provide a check and to give an indication of the values of δ^2y'' . These initial values will often be obtainable from a solution in series without requiring the evaluation of a large number of terms. In some cases it may be necessary to carry out a few steps of an integration at a small interval before starting the main integration.

Example: $y'' = (1-x^2)y$, $y(0) = 0$, $y'(0) = 1$, $\delta x = 0.1$. $y(0.1)$ and $y(0.2)$ evaluated from the series solution:

$$y = x + \tfrac{1}{6}x^3 - \tfrac{1}{24}x^5 + O(x^7).$$

| x | $1-x^2$ | y'' | $\delta^2 y''$ | | y | $\delta^2 y$ | $\delta^4 y$ (check) |
|-----|---------|-------|----------------|-----|---------------------|--------------------|-------------------------|
| 0.0 | 1.00 | 0 | | 0 | 0 | | |
| | | | 992 | -51 | | 10016 ₆ | |
| 0.1 | 0.99 | .0992 | 941 | -53 | .10016 ₆ | 10115 ₄ | 98 ₈ |
| | | | 837 | -57 | .20132 ₀ | 10307 ₈ | 93 ₆ |
| 0.2 | 0.96 | .1933 | | | | 192 ₄ | -10 ₃ |
| | | | 676 | -65 | .30439 ₈ | 275 ₇ | 83 ₃ |
| 0.3 | 0.91 | .2770 | | | | 10583 ₅ | -16 ₃ |
| | | | 450 | -71 | .41023 ₃ | 342 ₇ | 67 ₀ |
| 0.4 | 0.84 | .3446 | | | | 10926 ₂ | -22 ₆ |
| | | | 153 | | .51949 ₅ | 387 ₁ | 44 ₄ |
| 0.5 | 0.75 | .3896 | | | | 11313 ₃ | |
| 0.6 | 0.64 | .4049 | | | .63262 ₈ | | |

Here the numbers above the broken line are obtained from the series solution. In this example, y is an odd function of x , so that we have $\delta^2 y''(0) = 0$, as well as the value $\delta^2 y''(0.1) = -0.0051$. The value $\delta^2 y(0.1) = 0.000988$ is obtained in two ways, namely (a) from the first three values of y'' and the integration formula (7.2), and (b) from the first three values of y . Agreement between these values forms a check on the starting conditions, and also checks that the $\delta^2 y''$ term in the integration formula has been taken with the right sign; this term has the coefficient $+\frac{1}{12}$ here, whereas in the formula for a single integration, the first term in the correction to the trapezoidal formula has the coefficient $-\frac{1}{12}$. This makes it rather easy to make a mistake of sign at this point, and it is as well to have a check that the right sign has been taken.

If the integration has been taken to $x = 0.5$, the numbers above the full line will have been calculated; we will consider the integration through the next step, $x = 0.5$ to 0.6 . From the run of the third differences $\delta^3 y''$, the next value may be expected to be about -76 , giving $\delta^2 y''(0.5)$ about -302 ; a twelfth of this is -25 which gives

$$\delta^2 y(0.5) = (0.01)(0.3896 - 0.0025) = 0.003871.$$

This is entered in the $\delta^2 y$ column and checked by forming the value of $\delta^2[\delta^2 y(0.4)]$. By operating on both sides of (7.3) with δ^2 , we have

$$\delta^2(\delta^2 y)_0 = (\delta x)^2[\delta^2 y''_0 + \tfrac{1}{12}\delta^4 y''_0], \tag{7.4}$$

and use of this formula provides a good check on the values of $\delta^2 y$; it should be noted that the contribution from $\delta^4 y''_0$ may have to be included in (7.4) although it is negligible in (7.2).

The value of $y(0.6)$ is then built up from the value of $\delta^2 y(0.5)$, and from it the value of $y''(0.6)$ is obtained from the differential equation. This value is

$$y''(0.6) = 0.4049,$$

whence $\delta^2 y''(0.5) = -297$; the estimate was adequate and no recalculation of this interval is necessary. The calculation has reached the stage from which we started, only one interval further on, and a similar calculation for the next interval can now be undertaken.

The differences of the values of y'' form a check against random mistakes in these values, and use of formula (7.4) provides a current check of the values of $\delta^2 y$. The other process which needs checking is the twofold summation of these differences to give the solution y .

There are various ways of carrying out this check. If values of y have been built up by two successive summations, of δ^2y to δy and of δy to y , as illustrated in this example, a good check is provided by evaluating the second differences of y by the method of § 4.45, which does not involve the calculation of first differences, and verifying that the values so obtained reproduce the values of δ^2y . This check can be applied as each value of y is obtained, but is best carried out occasionally, say every ten intervals, in such a way as to verify the values obtained since the previous check.

Another check, which can only be carried out on a series of values of y , is provided by taking a set of alternate values of y and differencing them to second differences, taking the corresponding values of y'' and differencing as far as necessary for use in formula (7.2), and verifying that these values of δ^2y , and of y'' and its differences, do satisfy formula (7.2) with (δx) equal to twice the integration interval. It will usually be necessary to use higher orders of differences in y'' in this check than in the integration, but central differences of higher order than the second are available at this stage.

Example: To check the solution of $y'' = (1-x^2)y$ obtained above. Copying the values of y'' and y at intervals $\delta x = 0.2$, and differencing them, we have the second to sixth columns in the following table:

| x | y'' | δ^2y'' | δ^4y'' | y | δ^2y | $(\delta x)^2(y'' + \frac{1}{12}\delta^2y'' - \frac{1}{240}\delta^4y'')$ |
|-----|-------|---------------|---------------|---------------------|------------------|--|
| 0.0 | 0 | 0 | | 0 | | |
| .2 | .1933 | -420 | -70 | .20132 ₀ | 759 ₃ | $4(1933 - 35_0 - 0_3) = 759_1$ |
| .4 | .3446 | -910 | | .41023 ₃ | | |
| .6 | .4049 | | | | | |

The last column gives the calculation of $\delta^2y(0.2)$, for $\delta x = 0.2$, from the values of y'' and its differences. Agreement to a unit in the sixth decimal with the value of δ^2y is not to be expected; but this figure is only a guarding figure.

The interval δx used, and the number of figures kept in the different parts of the calculation, will depend on the equation, the data occurring in it, and the accuracy required in the results. This example is representative of the accuracy which it is convenient to keep in many calculations concerned with integration of equations occurring in scientific or technical problems. The last figure in y in this example is a guarding figure only, and could well be omitted if the function $f(x, y)$ in the equation involved some experimentally determined function which is not known to better than 1 part in 1000. In working to this accuracy, the first estimate of δ^2y'' in each interval, its division by 12, the addition of the result to y'' , can be done mentally; so can the multiplication by $(\delta x)^2$

when δx is a power of 10, as will often be the case. Then the first number written down is $(y'' + \frac{1}{12}\delta^2 y'')$ or $\delta^2 y$, and the value of $\delta^2 y$ is immediately checked by differencing. Such an integration can be carried out quite quickly.

The smaller the interval δx taken, the better the estimate of $\delta^2 y''$ can be made, and the smaller the quantity $(\delta x)^2$ by which this is multiplied. But it is not advisable to take very small intervals, first, because the amount of work required to cover a given range of x increases as the length of integration interval used decreases, and, secondly, because effects of rounding errors in the values of $\delta^2 y$ may accumulate rather rapidly in the double summation to give y . If a large number of small intervals are taken, additional guarding figures may have to be taken to ensure that the cumulative effects of rounding errors are negligible to the accuracy required in the final results, and this makes the amount of work involved increase rather more than proportionately to the number of intervals. The interval length δx should therefore be taken, roughly speaking, about as large as is compatible with ease in the practical numerical working of the integration. As a rough working rule it should be taken so that for only about one interval in five does the calculation for an interval have to be repeated. If many have to be repeated the interval should be halved.

If $\delta^2 y''$ is not too large, then a good approximation to $\delta^2 y_0$ is $(\delta x)^2 y_0''$, so that to this approximation, for a function satisfying equation (7.1),

$$y_1 = 2y_0 - y_{-1} + (\delta x)^2 f(x_0, y_0).$$

Thus for such a function we always have a good approximation to the function one interval ahead of where we know y'' . It is this feature which makes the numerical integration of such an equation such a straightforward process.

The procedure for two simultaneous equations

$$y'' + f(x, y, z) = 0, \quad z'' + g(x, y, z) = 0$$

or more, with all first derivatives absent, is similar.

7.21. Change of the interval of integration

It will not always be advisable to keep the same interval length throughout an integration. It may happen that the suitable interval length δx varies by a factor 10 or even 100 over the range of x to be covered; then the use, over the whole range, of the small interval necessary over part of it might make the calculation so long as to be almost impracticable.

As already emphasized in § 6.41, it is advisable at any change of interval length to take an overlap between the integrations carried out with the two different interval lengths. The most usual changes of interval length are by factors 2, $2\frac{1}{2}$, or $\frac{1}{2}$. To increase the interval length by a factor 2, all that is necessary is to take alternate values of y and the corresponding values of y'' , for two or three intervals before the point at which the change of interval length is to be made, difference them, and check the formula (7.2) for the function values and differences at this new interval, and continue as if these were intervals of the integration at the new interval. It may be necessary to keep an extra decimal in y'' to get the same accuracy in y .

To decrease the interval length by a factor 2, some interpolation is required, but only of the simplest kind, namely the 'half-way' interpolation considered in § 5.21. Suppose that integration has been carried out with intervals $\delta x = h$ up to $x = X$, and it is required to continue with intervals $\delta x = \frac{1}{2}h$. The integration with intervals $\delta x = h$ should be carried for one or two intervals beyond $x = X$, to give the central differences needed in the half-way interpolation. Then $y(X - \frac{1}{2}h)$ and $y''(X - \frac{1}{2}h)$ should be interpolated and $y''(X - \frac{1}{2}h)$ also calculated from the value of $y(X - \frac{1}{2}h)$ and the differential equation, to check. Then

$$\delta^2 y(X - \frac{1}{2}h) = y(X - h) - 2y(X - \frac{1}{2}h) + y(X)$$

should be calculated and compared with the value obtained from formula (7.2) at the smaller interval. This checks the interpolation of $y(X - \frac{1}{2}h)$; this check is most important since the whole subsequent integration would be vitiated by a mistake in this value. The integration then proceeds, starting from the values of $y(X - \frac{1}{2}h)$ and $y(X)$.

Example: To continue the integration of $y'' = (1 - x^2)y$ from $x = 0.5$, using intervals $(\delta x) = 0.05$.

The value of $\delta^4 y(0.5)$ for $(\delta x) = 0.1$ will be about -29_7 , so the interpolated value of $y(0.45)$ is

$$\begin{aligned} y(0.45) &= 0.41023_3 + \frac{1}{2}(10926_2) - \frac{1}{16}(342_7 + 387_1) + \frac{3}{256}(-22_6 - 29_7) \\ &= 0.41023_3 + 5463_1 - 45_{62} - 0_{61} = 0.46440_2. \end{aligned}$$

The interpolated value of $y''(0.45)$ is

$$\begin{aligned} y''(0.45) &= 0.3446 + \frac{1}{2}(450) - \frac{1}{16}(-226 - 297) \\ &= 0.3446 + 225 + 33 = 0.3704 \end{aligned}$$

(the fourth-difference contribution is negligible), whereas the value calculated from the differential equation is

$$y''(0.45) = [1 - (0.45)^2](0.46440_2) = 0.37036,$$

which, to four decimals, agrees with the interpolated value. Thus, starting from $x = 0.4$, we have the following values

| x | y'' | $\delta^2 y''$ | $y'' + \frac{1}{12} \delta^2 y''$ | y | $\delta^2 y$ |
|-----|-------|----------------|-----------------------------------|---------------------|-------------------|
| 0.4 | .3446 | | | .41023 ₃ | |
| | | 258 | | | 5416 ₉ |
| .45 | .3704 | -66 | .3698 ₅ | .46440 ₂ | 92 ₄ |
| | | 192 | | | 5509 ₃ |
| .5 | .3896 | | | .51949 ₅ | |

The value of $\delta^2 y(0.45)$ derived from the values of y'' is 92_{46} ; the difference between this value and the value 92_4 derived from the values of y is within the tolerance for the effect of the rounding error in the interpolated value. The integration can therefore be continued from these values.

Note: In the integration of this equation to this accuracy, it would not actually be necessary, or advisable, to decrease the interval of integration at this point; this case is only considered here as an example of the procedure.

The treatment of a change of interval length by a factor $2\frac{1}{2}$ is similar. Suppose, for example, the integration has been taken to $x = 0.30$ by intervals of 0.02 and it is desired to change to intervals of 0.05 . The values of y and y'' at $x = 0.25$ are obtained by half-way interpolation between the values at $x = 0.24$ and 0.26 , $y''(0.25)$ is checked as above, and $y(0.25)$ checked by verifying formula (7.2) at the larger interval.

7.22. Variants of the method

There are several variants of this method, some of which can be combined.

Instead of evaluating y'' for each value of x , forming $(y'' + \frac{1}{12} \delta^2 y'')$ and multiplying this by $(\delta x)^2$, we could evaluate $(\delta x)^2 y''$ and form $\delta^2 y$ as

$$\delta^2 y_0 = (\delta x)^2 y''_0 + \frac{1}{12} \delta^2 [(\delta x)^2 y'']_0.$$

At a change of interval length, the entries in the column of $(\delta x)^2 y''$ would be different for the two lengths of interval.

Another variant is as follows. If we operate on both sides of formula (7.2) with the repeated central sum operator $\sigma^2 = \delta^{-2}$, it becomes

$$y_0 = (\delta x)^2 [\sigma^2 y''_0 + \frac{1}{12} y''_0 - \frac{1}{240} \delta^2 y''_0]. \quad (7.5)$$

If this formula is used, the *aggregate* contributions from the $\frac{1}{12} \delta^2 y''_0 - \frac{1}{240} \delta^4 y''_0$ terms in (7.2) are evaluated separately for each interval, instead of being built up from contributions from successive intervals. This avoids accumulation of rounding and truncation errors in these contributions.

The process of starting the integration is rather more complicated, as initial values for the double sum $\sigma^2 y''$ have to be evaluated. The process does not avoid estimation, since if the integration has been carried to $x = x_0$, then for the end of the step from x_0 to x_1 we have

$$y_1 = (\delta x)^2 [\sigma^2 y''_1 + \frac{1}{12} y''_1 - \frac{1}{240} \delta^2 y''_1];$$

neither y_1 nor y_1'' is known at this stage, only the value of $\sigma^2 y_1''$ and the relation $y_1'' = f(x_1, y_1)$ between y_1 and y_1'' ; so this is an implicit equation for y_1 , and unless it happens to be linear the solution of it may well be more trouble than the integration process.

This variant requires special treatment at a point at which the interval length δx changes.

7.23. Numerov's method

If the equation to be solved is linear, say

$$y'' = f(x)y + g(x), \quad (7.6)$$

then the solution, to the accuracy given by neglecting the $\delta^4 y''$ term in (7.2), can be obtained without any estimation as follows. In (7.3)

$$\delta^2 y_0'' = y_1'' - 2y_0'' + y_{-1}'',$$

and if y'' is given by (7.6) this is

$$\delta^2 y_0'' = (f_1 y_1 - 2f_0 y_0 + f_{-1} y_{-1}) + \delta^2 g_0,$$

so that (7.3) can be written

$$\begin{aligned} [1 - \tfrac{1}{12}(\delta x)^2 f_1] y_1 - 2[1 - \tfrac{1}{12}(\delta x)^2 f_0] y_0 + [1 - \tfrac{1}{12}(\delta x)^2 f_{-1}] y_{-1} \\ = (\delta x)^2 [f_0 y_0 + g_0 + \tfrac{1}{12} \delta^2 g_0], \end{aligned}$$

$$\text{or} \quad \delta^2 \{ [1 - \tfrac{1}{12}(\delta x)^2 f] y \}_0 = (\delta x)^2 [f_0 y_0 + g_0 + \tfrac{1}{12} \delta^2 g_0]. \quad (7.7)$$

This treatment is usually ascribed to Numerov.[†] Written as a relation between three successive values of y , (7.7) is

$$[1 - \tfrac{1}{12}(\delta x)^2 f_1] y_1 = [2 + \tfrac{5}{6}(\delta x)^2 f_0] y_0 - [1 - \tfrac{1}{12}(\delta x)^2 f_{-1}] y_{-1} + (\delta x)^2 [g_0 + \tfrac{1}{12} \delta^2 g_0]. \quad (7.8)$$

A correction for the leading terms in the error of this formula can be evaluated by the following method, due to Olver.[‡] Since in this method y'' is never evaluated, it is convenient to express the corrections in terms of the differences of y itself. From formula (4.43) we have

$$(\delta x)^2 y_0'' = \delta^2 y_0 - \tfrac{1}{12} \delta^4 y_0 + \tfrac{1}{90} \delta^6 y_0 - \tfrac{1}{560} \delta^8 y_0 + O(\delta x)^{10}, \quad (7.9)$$

so that if y satisfies equation (7.6)

$$\delta^2 y = (\delta x)^2 [f(x)y + g(x)] + \tfrac{1}{12} \delta^4 y - \tfrac{1}{90} \delta^6 y + \tfrac{1}{560} \delta^8 y + O(\delta x)^{10}, \quad (7.10)$$

[†] B. Numerov, *Publ. de l'Observ. astrophysique central de Russie*, **2** (1933), 188; see also M. F. Manning and J. Millman, *Phys. Rev.* **53** (1938), 673, and for a similar method for a pair of simultaneous equations, M. V. Wilkes, *Proc. Camb. Phil. Soc.* **36** (1940), 204, and for a system of simultaneous equations with constant coefficients, D. R. Hartree, *Journ. Inst. Elect. Eng.*, vol. 103, Part B, supplement No. 1 (1956), 82.

[‡] For a similar treatment of a non-linear equation, see F. W. J. Olver, *Proc. Camb. Phil. Soc.* **46** (1950), 570, § 4.

and operation on both sides of this relation with $(1 + \frac{1}{12}\delta^2)$ gives

$$\delta^2[\{1 - \frac{1}{12}(\delta x)^2 f\}y] = (\delta x)^2[f(x)y + (g + \frac{1}{12}\delta^2 g)] - \frac{1}{240}\delta^6 y + \frac{13}{15120}\delta^8 y + O(\delta x)^{10}. \quad (7.11)$$

If the terms of sixth and higher orders in δx are neglected, this is equation (7.7). Let z be the solution of this approximate equation, with the same initial conditions as those specified for y ; that is

$$\delta^2[\{1 - \frac{1}{12}(\delta x)^2 f\}z] = (\delta x)^2 f(x)z + (g + \frac{1}{12}\delta^2 g), \quad (7.12)$$

and let the solution of (7.11) be

$$y = z + \eta.$$

Then, on subtracting (7.12) from (7.11), it follows that η satisfies

$$\delta^2[\{1 - \frac{1}{12}(\delta x)^2 f\}\eta] = (\delta x)^2 f(x)\eta - \frac{1}{240}\delta^6 y + \frac{13}{15120}\delta^8 y + O(\delta x)^{10}. \quad (7.13)$$

From equation (7.13) it follows that η is of order $(\delta x)^4$, hence $\delta^6 y$ differs from $\delta^6 z$ by terms of order $(\delta x)^{10}$. Hence (7.13) can be replaced, with an error of order $(\delta x)^{10}$, by

$$\delta^2[\{1 - \frac{1}{12}(\delta x)^2 f\}\eta] = (\delta x)^2 f(x)\eta - \frac{1}{240}\delta^6 z + \frac{13}{15120}\delta^8 z. \quad (7.14)$$

Then if z is calculated from formula (7.12) and η from formula (7.14), neither of which involve any estimation, the aggregate truncation error in $y = z + \eta$ is of order $(\delta x)^8$.

7.3. First-order differential equations

For a first-order equation the following method, when applicable, seems the most convenient. It is based on the use of the integration formula

$$y_1 - y_0 = \frac{1}{2}(\delta x)[(y'_0 + y'_1) - \frac{1}{6}(\delta x)\{\delta y''_0 - \frac{1}{60}\delta^3 y''_0\}], \quad (7.15)$$

expressing an integral in terms of the integrand and the *differences of its derivative* (see § 6.22). It is applicable if the function f in the equation

$$y' = f(x, y)$$

is either given by an analytical formula, so that there is no difficulty in evaluating

$$y'' = \frac{\partial f}{\partial x} + \frac{\partial f}{\partial y} \frac{dy}{dx} = \frac{\partial f}{\partial x} + f \frac{\partial f}{\partial y}$$

to any accuracy required, or if $\partial f/\partial x$ and $\partial f/\partial y$ can be determined by numerical differentiation to the accuracy required in using the formulae of the method.

Two advantages of formula (7.15) have already been pointed out in § 6.22, namely, the small coefficient of the $\delta^3 y''$ term in the square bracket,

and the fact that if this term (and higher terms) are neglected, so that the formula becomes

$$y_1 - y_0 = \frac{1}{2}(\delta x)[(y'_0 + y'_1) - \frac{1}{6}(\delta x)(y''_1 - y''_0)], \quad (7.16)$$

it does not involve the values of any quantities at points outside the interval through which the integration is being taken.

Further, the values of y'' calculated for use in the integration formula can be used to give approximations to the successive values of y by use of formula (7.3), namely

$$\delta^2 y_0 = (\delta x)^2 [y''_0 + \frac{1}{12} \delta^2 y''_0]$$

with the term in $\delta^2 y''_0$ either omitted or estimated.

If the numbers occurring in the integration are arranged with those referring to the same value of x in a *column*, instead of in a row, the work can be arranged as in the example on p. 145.

Example:

$$y' = 1 - 2xy; \quad y = 0 \text{ at } x = 0 \quad (\text{compare } \S 6.43, \text{ equation (6.35)}).$$

In this case $y'' = -2xy' - 2y$. For this equation, it would be possible to use a series to start the integration, but here this will not be used, so as to show the procedure when the use of a series is not convenient.

Since in this equation the values of $2y$ occur in the formula for y' and y'' , it is convenient to accumulate values of $2y$ rather than y .

For the first interval the values of y_0 , y'_0 , and y''_0 are available and a first estimate of the value of $2y(0.1)$ is obtained by using these values in the first three terms of a Taylor series; this gives $2y(0.1) = 0.20$. From this, approximate values of $y'(0.1)$ and $y''(0.1)$ are found, and integration carried through the interval $x = 0$ to 0.1 , giving a better value of $2y(0.1)$, namely 0.1987 , from which better values of y' and y'' at the end of the first interval and a better final value of y are obtained. Use of this better value of y does not change the values of y' and y'' to the accuracy to which they are being used in the integration, and the revised integration through the first interval can then be taken as the first step in the main integration.

Suppose now that the integration has reached $x = 0.3$, so that the quantities to the left of the heavy line are known. The run of the second differences of $\delta^2 y''$ suggests that the value at 0.3 will be about 80 , so that

$$\delta^2[2y(0.3)] = (0.01)2[-1.063 + 0.007] = -0.02112.$$

This is written in on the last line but two, and the approximate value of $2y(0.4) = 0.71989$ is built up from it. The values $y'(0.4)$, $y''(0.4)$ are calculated from this value of $2y$, and then the integration carried out, giving $2y(0.4) = 0.71987$. The difference of this value from the trial value 0.71989 is not such as to affect y' or y'' to the accuracy to which they are kept. The integration has now been taken to $x = 0.4$, and the sequence of operations can be repeated for the next interval. *Notes:* (i) Since with the interval taken the values of $\delta y''$ are divided by 60 before being added to those of y' , it is adequate to keep y'' to one decimal fewer than y' .

(ii) It would be practicable to keep another decimal in $2y$ without using a smaller interval δx .

| x | 0.0 | 0.1 | 0.2 | 0.3 | 0.4 |
|---|-------|--------|--------|--------|--------|
| $-2xy$ | 0 | -.020 | -.0779 | -.1696 | -.2880 |
| $y' = 1-2xy$ | 1 | +.980 | .9221 | .8304 | .7120 |
| $-2xy'$ | 0 | -.196 | -.3688 | -.4982 | -.5696 |
| $-2y$ | 0 | -.200 | -.3894 | -.5652 | -.7199 |
| Sum = y'' | 0 | -.396 | -.758 | -1.063 | -1.289 |
| | | -395 | -363 | -305 | -226 |
| | | 32 | 58 | 79 | |
| $y'_0 + y'_1$ | 1.980 | 1.9801 | 1.9022 | 1.7525 | 1.5424 |
| $-\frac{1}{8}(\delta x)(y'_1 - y'_0)$ | +.007 | +.0066 | 60 | 51 | 38 |
| Sum = S | 1.987 | 1.9867 | 1.9082 | 1.7576 | 1.5462 |
| $S\delta x = \delta(2y)$ | .1987 | .19867 | .19082 | .17576 | .15462 |
| $2y$ | 0 | .1987 | .38949 | .56525 | .71987 |
| $\frac{2(\delta x)^2(y'' + \frac{1}{12}\delta^2 y'')}{\text{Next } \delta(2y)}$ | | -.790 | -.1506 | -.2112 | |
| Est. $\delta(2y)$ | | .19077 | .17576 | .15464 | |
| $\frac{2(\delta x)^2(y'' + \frac{1}{12}\delta^2 y'')}{\text{Next } 2y}$ | 0.20 | .38944 | .56525 | .71989 | |

7.31. Another method for a first-order equation

If in the equation in the form $y' = f(x, y)$, the function f is not such that y'' can be obtained to an accuracy adequate for use in the method of the preceding section, a formula involving only y' and its differences can be used. The appropriate central-difference formula, with an error term of the same order as that of (7.16), is

$$y_1 - y_0 = \frac{1}{2}(\delta x)[y'_0 + y'_1 - \frac{1}{12}(\delta^2 y'_0 + \delta^2 y'_1)]. \quad (7.17)$$

But this is not so convenient, because $\delta^2 y'_1$ involves y'_2 , which is not known until the *next* interval, from y_1 to y_2 , has been completed, and because the error term is larger than that of formula (7.16).

When the integration has reached x_0 , only y_0 , y'_0 , and backward differences from y_0 are available. There is an integration formula using backward differences from the beginning of the interval, but it is not satisfactory for practical use since the coefficients of the neglected higher-order differences are so large; that of the fourth difference is $\frac{251}{720}$ instead of $\frac{11}{720}$ for formula (7.16). Even if the trapezoidal approximation is used, it is necessary to estimate y_1 to give y'_1 , and to adjust this estimate until it agrees with the result of integration with the corresponding value of y'_1 ; and without the value of y'_0 it is more difficult to obtain a good approximation for the first estimate. Even when y'_1 is known, $\delta^2 y'_1$ also has to be estimated, and this estimate cannot be confirmed until the *next* interval has been completed. An examination of the differences of y' and of y in the worked example of the previous section will show the advantage of having values of y'' available so that the only estimate that has to be made is one of $\delta^2 y''$.

7.32. First-order linear equations

With a first-order linear equation

$$y' + f(x)y = g(x) \quad (7.18)$$

there is a strong temptation to follow the standard method of textbooks on differential equations and reduce the solution of the equation to a quadrature by use of an integrating factor. There are occasions for which this is useful (see § 7.61 for an example). But the writer's experience has been that in the great majority of cases arising for solution in practice, this is a temptation to be resisted, and that it is considerably easier to evaluate the solution by numerical integration of the equation as it stands than to evaluate the integrals in the solution in quadratures. An example has been given in § 1.1.

The reason for this can be illustrated by an example. Consider the equation

$$y' + 2xy = g(x) \quad (7.19)$$

with the condition $y = 0$ at $x = 0$, and with $g(x)$ a positive function whose maximum is of the order of unity, and which is given to six decimals, and is appreciable to this accuracy for $x < 20$ and negligible for higher values of x [$g(x) = x^2 e^{-x}$ for example]. The solution y has a maximum of order of magnitude unity, and tends to zero as $x \rightarrow \infty$. The integrating factor of equation (7.19) is e^{x^2} , and over the range of x where $g(x)$ is appreciable, e^{x^2} increases by a factor of the order of 10^{170} . The evaluation of the integral $\int_0^x e^{x^2} g(x) dx$, when the numerical magnitude of the integrating factor covers such a wide range, offers numerical difficulties (this integral might not even converge for large x , but the six-decimal accuracy to which $g(x)$ is supposed given would provide no data for evaluating it beyond $x = 20$); and even in the middle of this range, in the neighbourhood of $x = 10$, the solution would be calculated as the product of two numbers of the order of 10^{40} and 10^{-40} . These difficulties do not mean that the calculation could not be carried out in this way, but they do strongly suggest that it should not be carried out in this way in practice.

Fox and Goodwin† have given a process for numerical integration of equation (7.18) in which the linear character of the equation is used to avoid the estimation of y mentioned in § 7.31, in much the same way that Numerov's method for a second-order linear equation (7.6) uses the linear character of the equation to avoid estimation of $\delta^2 y$.

For integration through one interval, the trapezium rule with correction is

$$y_1 - y_0 = \frac{1}{2}(\delta x)(y'_0 + y'_1) + C.$$

Substitution for y' from the differential equation (7.18) and rearrangement of terms gives

$$[1 + \frac{1}{2}(\delta x)f_1]y_1 = [1 - \frac{1}{2}(\delta x)f_0]y_0 + \frac{1}{2}(\delta x)(g_0 + g_1) + C. \quad (7.20)$$

If z is the solution of the relation obtained from (7.20) by omitting the correction C to the trapezium rule, that is

$$[1 + \frac{1}{2}(\delta x)f_1]z_1 = [1 - \frac{1}{2}(\delta x)f_0]z_0 + \frac{1}{2}(\delta x)(g_0 + g_1), \quad (7.21)$$

and

$$y = z + \eta, \quad (7.22)$$

then

$$[1 + \frac{1}{2}(\delta x)f_1]\eta_1 = [1 - \frac{1}{2}(\delta x)f_0]\eta_0 + C. \quad (7.23)$$

Fox and Goodwin suggest first solving (7.21), and then solving (7.23) by an iterative procedure, the correction C being obtained from the values of y given by (7.21), (7.22) with the values of η obtained from the previous iteration, starting with $y = z$.

Another procedure is to integrate from y_0 to y_2 through two intervals $\delta x = h$ and through one interval $\delta x = 2h$, by means of the trapezium-

† L. Fox and E. T. Goodwin, *Proc. Camb. Phil. Soc.* **45** (1949), 373.

rule formula (7.21), eliminate the leading term in the truncation error at x_2 by Richardson's ' h^2 -extrapolation' process (see § 7.51; equation (7.36)), and continue the integration from the corrected value so obtained. If, starting from $z_0 = y_0$ at $x = x_0$, the results at $x = x_2$ of using formula (7.21) with two intervals $\delta x = h$ and with one interval $\delta x = 2h$ are z_2 and z_2^* respectively, then

$$y_2 = z_2 + \frac{1}{3}(z_2 - z_2^*) + O(\delta x)^5 \quad (7.24)$$

and
$$y_1 = z_1 + \frac{1}{6}(z_2 - z_2^*) + O(\delta x)^4. \quad (7.25)$$

The errors $O(\delta x)^5$ in formula (7.24) accumulate over a set of successive pairs of intervals, so that the aggregate error over a given range X of x is $O(\delta x)^4$; the errors of formula (7.25) do not accumulate, so that the lower order of the error in this formula can be tolerated.

7.33. Second-order equation with the first derivative present

The most convenient practical treatment of a second-order equation with the first derivative present depends on the form of the equation. For a linear equation

$$y'' + f(x)y' + g(x)y = h(x) \quad (7.26)$$

the term in y' can be eliminated by the use of

$$Y = y \exp\left[\frac{1}{2} \int f(x) dx\right];$$

this gives

$$Y'' + [g(x) - \frac{1}{2}f'(x) - \frac{1}{4}\{f(x)\}^2]Y = h(x)\exp\left[\frac{1}{2} \int f(x) dx\right] \quad (7.27)$$

which reduces the equation to the form treated in § 7.2. This is likely to be a convenient reduction for the homogeneous equation, in which $h(x) = 0$. If $h(x)$ is not zero, the exponential factor may make the right-hand side of (7.27) vary too rapidly to be convenient for numerical work; though since Y has to be divided by a corresponding exponential factor to give the solution y required, it may be possible to drop the less significant digits of Y as the solution proceeds.

If the equation is linear in y' , though not in y :

$$y'' + f(x)y' + g(x, y) = 0, \quad (7.28)$$

the term in y' can be eliminated by the same change of variable, though the resulting equation is not so convenient as (7.27). If, however, $g(x, y)$ is periodic in x and a periodic solution of y is required, this reduction of the equation is not very convenient, as the function Y will not in general be periodic. Then it is probably best to use the equation in the form (7.28).

A general method of treating the general second-order equation

$$y'' + f(x, y, y') = 0 \quad (7.29)$$

is to regard it as two simultaneous first-order equations

$$y' = z, \quad z' + f(x, y, z) = 0,$$

the latter being integrated first in each interval. That is, y'' is first integrated to give y' , and then y' is integrated to give y . The value of y at the end of the interval can be estimated by use of

$$\delta^2 y_0 = (\delta x)^2 [y''_0 + \frac{1}{12} \delta^2 y''_0],$$

of which only the term $\frac{1}{12} \delta^2 y''_0$ has to be estimated, and for the integration of y' to give y , y''_1 is known so that the integration formula (7.16) can be used for this integration. If the function $f(x, y, y')$ in equation (7.29) is such that y''' can be evaluated to adequate accuracy from values of x , y , and y' , then the method of § 7.3 can also be used for the integration of y'' to give y' . For example, for the van der Pol equation

$$y'' - (1 - y^2)y' + ky = 0,$$

we have

$$y''' - (1 - y^2)y'' + 2y(y')^2 + ky' = 0,$$

and can use the method of § 7.3 twice in each interval, once to integrate y'' to give y' and then to integrate y' to give y . This reduction of a second-order equation to two first-order equations should *not* generally be used for a second-order equation with the first derivative absent.

7.34. Equations of order higher than the second

If it is required to treat numerically an equation of order higher than the second, it is best to break down the integration through each interval into a sequence of single and twofold integrations. In each interval the highest derivative should be integrated first, and the lower-order derivatives in succession; then, apart perhaps from the integration of the highest-order derivatives, formula (7.16) can be used for any single integration required.

7.4. Taylor series method†

There is another method, which is in principle applicable to equations, of suitable form, of any order. Its limitation is that it is only suitable for equations in which the relation between the derivatives is given by an analytical formula, so that it can be differentiated formally as many

† See, for example, J. C. P. Miller, *British Association Mathematical Tables*, Part-volume B, *The Airy Integral* (1946), Introduction, § 5.

times as is required. For example, in the case of the equation

$$y' = x^2 - y^2 \quad (7.30)$$

we have in succession

$$\left. \begin{aligned} y'' &= 2(x - yy'), & y''' &= 2[1 - \{yy'' + (y')^2\}] \\ y^{iv} &= -2[yy''' + 3y'y''], & y^v &= -2[yy^{iv} + 4y'y''' + 3(y'')^2] \text{ etc.} \end{aligned} \right\} \quad (7.31)$$

It would be possible here to substitute for y' from (7.30) in the first of equations (7.31) before differentiating, but this would lead to more complicated formulae, and it is better to carry out the substitution numerically rather than algebraically.

Consider first a first-order equation such as (7.30), and suppose that the solution has been taken to $x = x_0$, so that y_0 is known. Then $y_0'', y_0''', y_0^{iv}, \dots$ can be calculated in turn from a set of relations such as (7.31), and then y_1 can be calculated from the Taylor series

$$y_1 = y(x_0 + \delta x) = y_0 + (\delta x)y_0' + \frac{1}{2!}(\delta x)^2 y_0'' + \frac{1}{3!}(\delta x)^3 y_0''' + \dots$$

It is convenient to arrange the numerical work so that the terms containing odd powers of δx and those containing even powers are added up separately:

$$\left. \begin{aligned} S_{\text{even}} &= y_0 + \frac{1}{2!}(\delta x)^2 y_0'' + \frac{1}{4!}(\delta x)^4 y_0^{iv} + \frac{1}{6!}(\delta x)^6 y_0^{vi} + \dots \\ S_{\text{odd}} &= (\delta x)y_0' + \frac{1}{3!}(\delta x)^3 y_0''' + \frac{1}{5!}(\delta x)^5 y_0^v + \dots \end{aligned} \right\} \quad (7.32)$$

$$\left. \begin{aligned} \text{Then} & & y_1 &= S_{\text{even}} + S_{\text{odd}} \\ \text{and} & & y_{-1} &= S_{\text{even}} - S_{\text{odd}} \end{aligned} \right\} \quad (7.33)$$

This calculation of y_{-1} , the starting-point for the previous interval, from y and its derivatives at $x = x_0$ is a very good check; y_0 has been calculated from y and its derivatives at x_{-1} , so that almost all the numbers involved in the calculation of y_{-1} from y_0 by (7.33) are different from those involved in the original calculation of y_0 from y_{-1} .

There is no particular reason for working with the derivatives themselves rather than with convenient multiples of them. In this case the convenient multiples are the quantities $Y^{(n)}$ defined by

$$Y^{(n)} = \frac{1}{n!}(\delta x)^n y^{(n)};$$

these are sometimes called 'reduced derivatives'. That is,

$$Y^{(0)} = y, \quad Y^{(1)} = (\delta x)y', \quad Y^{(2)} = \frac{1}{2}(\delta x)^2 y'', \quad \dots$$

Then, for example, (7.30), (7.31) become

$$\begin{aligned} Y^{(1)} &= (\delta x)[x^2 - \{Y^{(0)}\}^2], \\ Y^{(2)} &= (\delta x)[x \delta x - Y^{(0)}Y^{(1)}], \\ Y^{(3)} &= \frac{1}{3}(\delta x)[(\delta x)^2 - 2Y^{(0)}Y^{(2)} - \{Y^{(1)}\}^2], \\ Y^{(4)} &= -\frac{1}{2}(\delta x)[Y^{(0)}Y^{(3)} + Y^{(1)}Y^{(2)}] \quad \text{etc.}, \end{aligned}$$

and (7.32) becomes

$$\left. \begin{aligned} S_{\text{even}} &= Y_0^{(0)} + Y_0^{(2)} + Y_0^{(4)} + Y_0^{(6)} + \dots \\ S_{\text{odd}} &= Y_0^{(1)} + Y_0^{(3)} + Y_0^{(5)} + \dots \end{aligned} \right\}. \quad (7.34)$$

No special procedure is necessary for starting the integration.

For a second-order equation y_1' has to be calculated from a Taylor series as well as y_1 . We have

$$y'_{\pm 1} = y'_0 \pm (\delta x)y''_0 + \frac{1}{2!}(\delta x)^2 y'''_0 \pm \frac{1}{3!}(\delta x)^3 y^{(4)}_0 + \dots,$$

and hence

$$Y_{\pm 1}^{(1)} = (\delta x)y'_{\pm 1} = Y_0^{(1)} \pm 2Y_0^{(2)} + 3Y_0^{(3)} \pm 4Y_0^{(4)} + \dots;$$

so that if we write

$$S'_{\text{even}} = 2Y_0^{(2)} + 4Y_0^{(4)} + 6Y_0^{(6)} + \dots, \quad S'_{\text{odd}} = Y_0^{(1)} + 3Y_0^{(3)} + 5Y_0^{(5)} + \dots,$$

the reduced first derivative at $x = x_1$ is

$$Y_0^{(1)} = S'_{\text{even}} + S'_{\text{odd}}$$

and the check on the integration is provided by

$$Y_{-1}^{(1)} = S'_{\text{odd}} - S'_{\text{even}}.$$

By taking the series (7.34) to several terms, it is practicable to make the truncation error of considerably higher order in (δx) than it is in the case of formula (7.15) or (7.17), and so to work with a larger interval δx or alternatively to a greater number of significant figures. Results to a large number of figures will probably not be required except for equations which do satisfy the conditions for this method to be practicable, and in such cases it is a very powerful method.

7.5. Other procedures

A number of other procedures have been proposed for the numerical integration of differential equations. A few will be summarized in the following sections.

7.51. Richardson's 'deferred approach to the limit'

In most of the procedures so far explained it has been the purpose to make each interval of the integration correct, within the tolerance for

rounding error, before going on to the next. This is done by keeping the truncation error in each interval less than the rounding error. An alternative procedure is to carry out a whole integration using a very simple integration formula for which the truncation error is greater than the rounding error, and only correcting for the truncation error after the whole integration is completed. Such a process has been called by L. F. Richardson† a ‘deferred approach to the limit’; a process of this kind is applicable to quadrature as well as to the numerical integration of differential equations.

If in integrating the first-order equation

$$y' = f(x, y)$$

we use simply the trapezoidal formula

$$\delta y_{\frac{1}{2}} = y_1 - y_0 = \frac{1}{2}(\delta x)(f_0 + f_1), \quad (7.35)$$

the result y at a given value of x will depend on the interval length δx used in the integration as well as on x . Let us express this by writing this result as $y(x, \delta x)$; the solution of the differential equation is the limit of this as $\delta x \rightarrow 0$, namely $y(x, 0)$.

Now in each interval the error in δy calculated by (7.35) is of order $(\delta x)^3$. The number of intervals required to cover a given range of x is inversely proportional to δx ; hence the aggregate truncation error is of order $(\delta x)^2$. Such an error in y results in an error in y' of order $(\delta x)^2$, which makes an additional error of order $(\delta x)^3$ in each δy , which is of the same order as the truncation error in that interval alone. Thus the aggregate error at any given x is of order $(\delta x)^2$.

If now two separate integrations are carried out, using the same integration formula (7.35), with different interval lengths (δx) , then the leading term in the aggregate truncation error can be eliminated by extrapolating to $\delta x = 0$, linearly in $(\delta x)^2$, at each value of x . The most convenient way of doing this in practice is by use of one set of intervals $\delta x = h$ and another set $\delta x = 2h$; the extrapolation process is represented graphically in Fig. 12. A convenient numerical process is represented by the formula

$$y(x, 0) = y(x, h) - \frac{1}{3}[y(x, 2h) - y(x, h)]. \quad (7.36)$$

This process has been called by Richardson ‘ h^2 -extrapolation’. It is important to ensure that cumulative rounding errors do not vitiate this extrapolation to $\delta x = 0$.

It is in principle possible to carry out this process of extrapolation to

† L. F. Richardson, *Phil. Trans. Roy. Soc.* **226** (1927), 300.

$\delta x = 0$ from results calculated for more than two different interval lengths, but this is not a satisfactory procedure in many cases.

If the aggregate error over a given range of x is $O(\delta x)^4$, as with Numerov's method (§ 7.23), a similar process of extrapolation to $\delta x = 0$, linearly in $(\delta x)^2$, can be used.

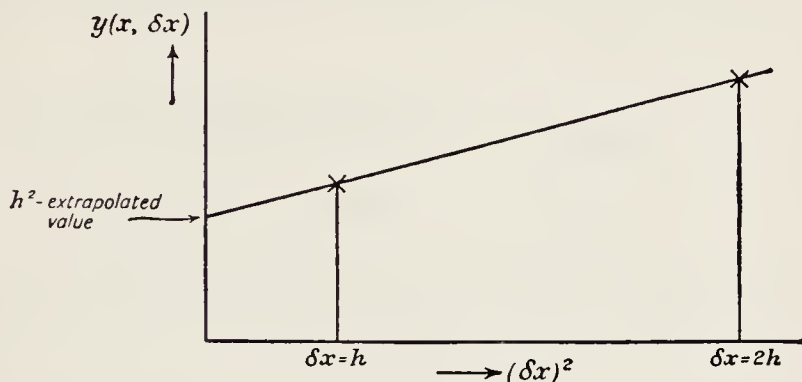


FIG. 12.

7.52. Iterative processes

The solution of the equation

$$y' = f(x, y), \quad y(x_0) = y_0$$

can formally be obtained by constructing a sequence of functions $y_{(n)}(x)$ by

$$y_{(n+1)}(x) = y_0 + \int_{x_0}^x f\{x, y_{(n)}(x)\} dx. \quad (7.37)$$

If the integral is evaluated by quadrature, this process of iterative quadrature is the numerical equivalent of Picard's process in the formal analytical theory of differential equations. It is sometimes useful for starting an integration, but unless a good approximation is available to use as a first approximation $y_0(x)$ in the right-hand side of (7.37), it is seldom useful for carrying the integration of an equation with one-point boundary conditions over a range of x , because the convergence of the successive functions $y_{(n)}$ to the solution of the equation is too slow. A form of iterative quadrature may, however, be useful in dealing with problems with two-point boundary conditions (see § 7.61).

As an example of another, more practical, kind of iterative process, consider the equation

$$y'' = f(x, y). \quad (7.38)$$

A sequence of functions $y_{(n)}$ can be formed by

$$\delta^2 y_{(n)} = (\delta x)^2 [f\{x, y_{(n)}(x)\} + \frac{1}{12} \delta^2 y_{(n-1)}'' - \frac{1}{240} \delta^4 y_{(n-1)}'']. \quad (7.39)$$

Here only the leading term on the right-hand side depends on the

function $y_{(n)}$ currently being evaluated; the 'correcting' terms, involving $\delta^2 y''$, $\delta^4 y''$ are derived from the previously calculated function $y_{(n-1)}$. This avoids any estimation of $\delta^2 y''$, and enables $\delta^4 y''$ and $\delta^6 y''$, and higher differences, to be included.

A different iterative process is obtained if y'' on the left-hand side of (7.38) is expressed in terms of y and its differences. Substitution for y'' in formula (7.9) gives

$$\delta^2 y = (\delta x)^2 f(x, y) + \frac{1}{12} \delta^4 y - \frac{1}{90} \delta^6 y + \frac{1}{560} \delta^8 y + O(\delta x)^{10}. \quad (7.40)$$

Then a sequence of functions $y_{(n)}$ can be formed by means of the iterative formula

$$y_{(n+1)} = (\delta x)^2 [\sigma^2 f\{x, y_{(n+1)}\} + \frac{1}{12} \delta^2 y_{(n)} - \frac{1}{90} \delta^4 y_{(n)}] + O(\delta x)^8, \quad (7.41)$$

where σ^2 is the twofold sum operator as in § 7.22. Here the 'correcting' terms in the evaluation of $y_{(n+1)}$ are expressed in terms of the differences of the previous function $y_{(n)}$ itself, instead of in terms of the differences of its second derivative. A similar treatment can be applied to first-order equations, and various examples of its application have been given by Fox and Goodwin.†

7.53. The Madelung transformation

For evaluating oscillatory solutions of a second-order homogeneous linear equation with the first derivative absent:

$$y'' + f(x)y = 0, \quad (7.42)$$

a transformation due to Madelung‡ is sometimes useful.

$$\text{Let us write} \quad y = F(x) \exp\left[i \int g(x) dx\right], \quad (7.43)$$

where F and g are to be real; this is equivalent to describing the oscillating function y at each point in terms of a local amplitude $F(x)$ and local phase $\phi(x) = \int g(x) dx$. The real and imaginary parts of (7.43) clearly give two linearly independent solutions of (7.42). Substitution of (7.43) into (7.42) and separation of real and imaginary parts gives

$$F'' - Fg^2 + fF = 0, \quad (7.44)$$

$$2F'g + Fg' = 0. \quad (7.45)$$

The second of these two equations is integrable and gives

$$F^2 g = \text{const.} = A \quad (\text{say}), \quad (7.46)$$

and substitution in (7.42) then gives

$$F'' - A^2/F^3 + fF = 0. \quad (7.47)$$

† L. Fox and E. T. Goodwin, *Proc. Camb. Phil. Soc.* **45** (1949), 373.

‡ E. Madelung, *Zeit. für Phys.* **67** (1931), 516.

Thus the evaluation of the two functions F and g can be separated, equation (7.47) being first solved for F , and g then determined.

Equation (7.47) is not linear, but it may be easier to integrate than the original equation (7.42), since the rapidly oscillating part of y has been taken out by the factor $\exp[i \int g(x) dx]$, and the function F describing the local amplitude of this oscillation will often vary relatively slowly.†

7.54. The Riccati transformation

For equation (7.42) with $f(x)$ negative, the Riccati transformation

$$\eta = y'/y = d(\log y)/dx, \quad \eta' + \eta^2 + f(x) = 0 \quad (7.48)$$

is sometimes useful, especially if $f(x)$ is negative over a considerable range of x . If $f(x)$ is negative and slowly varying, and $|f(x)|$ is large, a solution of equation (7.48) can sometimes be obtained by writing it in the form

$$\eta = \pm[-f(x) - \eta']^{\frac{1}{2}} \quad (7.49)$$

and solving this by iteration. The first approximation is

$$\eta = [-f(x)]^{\frac{1}{2}};$$

numerical differentiation then gives an approximation to η' in (7.49) and hence a better value of η . This is one of the few situations in which numerical differentiation may be useful as a tool in a practical numerical process.

7.6. Two-point boundary conditions

As an example of the treatment of two-point boundary conditions, consider the solution of

$$y'' = f(x, y)$$

subject to the conditions

$$y = y_0 \quad \text{at } x = x_0, \quad y = y_b \quad \text{at } x = b.$$

A step-by-step integration, starting from $x = x_0$, has to start from definite numerical values of y at the beginning and end of the first interval, and the result at any later value of x is determined by these two values of y . The former is given to be y_0 , but the latter, $y_1 = y(x_0 + \delta x)$, is not specified. If the variation of the solution $y(b)$ at $x = b$ with the value of y_1 is not too rapid, the following process can be used to find the solution satisfying the condition $x = b$. A set of integrations is carried out for a set of trial values of y_1 , and the value of $y(b)$ obtained as a function of y_1 . Interpolation (graphical or numerical) can then be used to obtain a

† For an example of the use of this transformation, see D. R. Hartree, R. L. Kronig, and H. Pedersen, *Physica*, **1** (1934), 895.

close approximation, say η_1 , to the value of y_1 which gives a solution for which $y(b)$ has the required value y_b . A further one or two integrations with values of y_1 in the immediate neighbourhood of η_1 then enables the solution satisfying the given condition at $x = b$ to be obtained by linear interpolation.

If the value of $y(b)$ is very sensitive to the value of y_1 such a process is not practicable in this simple form. Consider the variation of y at a fixed x with the trial value of y_1 , this is $\partial y(x)/\partial y_1$. This is a function of x , and, for a non-linear differential equation, corresponds to a 'complementary function' of a linear differential equation. It might behave approximately as e^{x^2} , in which case it would increase by a factor of about 10^{10} over a range of x from 0 to 5. This would mean that unless the choice of y_1 were correct to 0.00001, all trial solutions at $x = b = 5$ would have values of $|y(b)|$ of the order of 10^5 or larger. If the condition at $x = b$ were $y(b) = 1$, interpolation between two trial solutions with values of y_1 differing by 0.00001 would determine y_1 closely but would not determine the solution at all well except near $x = 0$. Further, the effects of rounding errors in the early intervals of the integration build up in much the same way as the function $\partial y(x)/\partial y_1$ so that it would be necessary to keep a large number of guarding figures, and probably to work to 15 or 20 decimals.

Such a situation is not rare; in the writer's experience, it is more likely than not to occur in equations with two-point boundary conditions which arise in real problems (as distinct from those which are made up to serve as textbook examples). In such a situation, however, a procedure of the same kind can be used, proceeding by stages in the x direction. Two solutions with different values of y_1 are carried to such a value of x , say x_a , that their behaviour indicates clearly enough whether the required solution lies between them or not; if not, other solutions are evaluated until a pair is found between which the required solution does lie. Let these be y_I and y_{II} , with values $(y_I)_1$ and $(y_{II})_1$ at the end of the first interval.

From the behaviour of the solutions y_I and y_{II} at $x = x_a$ and the expected behaviour of the solution required, an estimate is made of the fraction p of the difference between the functions y_I and y_{II} such that $y_I + p(y_{II} - y_I)$ is a fair approximation to the required solution. Linear interpolation will probably not be valid at $x = x_a$ but should be good enough to give one decimal in p , which is all that is wanted. Another solution, y_{III} , is then started, not from $x = x_a$ but from some smaller value x_1 of x at which linear interpolation between the solutions y_I and

y_{II} is valid to the accuracy to which the calculation is carried. Whether linear interpolation is valid can usually be tested by comparing (a) the value of $y''(x_1)$ interpolated linearly between the values for solutions y_I and y_{II} , that is

$$y_I''(x_1) + p[y_{II}''(x_1) - y_I''(x_1)],$$

and (b) the value of $y''(x_1)$ calculated from the interpolated value of $y(x_1)$. If the difference between these two values of $y''(x_1)$ is not enough to affect the last digit of $y(x_1 + \delta x)$, then, in the usual contexts in which this procedure is required, the linear interpolation is adequate.

Depending on the behaviour of y_{III} , either another solution is started from x_i , or a solution is started by linear interpolation between y_{III} and y_{II} or y_I at a point x_{ii} farther out, by a repetition of the process for selecting and starting the evaluation of the solution y_{III} . This process may have to be repeated several times before the value $x = b$ is reached.

7.61. Iterative quadrature

In the process considered in the previous section the solution satisfying the two-point boundary conditions is reached by evaluating a sequence of functions each of which *does* satisfy the differential equation but *does not* satisfy *all* the boundary conditions. An alternative procedure in some cases is to approach the solution required through a sequence of functions each of which satisfies *all* the boundary conditions, but does *not* satisfy the equation.

Consider, for example, the equation

$$y''' = -(1 + y^2)y'' \quad (7.50)$$

with boundary conditions

$$y = 0, \quad y' = 0 \quad \text{at } x = 0, \quad y' \rightarrow 1 \quad \text{as } x \rightarrow \infty. \quad (7.51)$$

Let $y_{(n)}(x)$ be a sequence of functions defined by

$$y_{(n+1)}''' = -(1 + y_{(n)}^2)y_{(n+1)}''. \quad (7.52)$$

If at any stage of the work $y_{(n)}$ is a known function of x , this is an equation for the next function of the sequence, namely $y_{(n+1)}$; it is linear and homogeneous in this unknown function and there is no difficulty in obtaining a solution of (7.52) satisfying *all three* of the boundary conditions (7.51). One integration gives

$$y_{(n+1)}'' = A \exp \left[- \int_0^x (1 + y_{(n)}^2) dx \right],$$

where A is, so far, an undetermined integration constant. Another integration gives

$$y'_{(n+1)} = A \int_0^\infty \exp\left\{-\int_0^x (1+y_{(n)}^2) dx\right\} dx;$$

the condition $y'(0) = 0$ has been satisfied by choosing the lower limit of the integral; the condition $y'(\infty) = 1$ can now be satisfied by choice of A , and gives

$$y'_{(n+1)} = \frac{\int_0^\infty \exp\left\{-\int_0^x (1+y_{(n)}^2) dx\right\} dx}{\int_0^\infty \exp\left\{-\int_0^x (1+y_{(n)}^2) dx\right\} dx},$$

and another integration from lower limit $x = 0$ satisfies the condition on $y(0)$, giving

$$y_{(n+1)} = \frac{\int_0^\infty \int_0^\infty \exp\left\{-\int_0^x (1+y_{(n)}^2) dx\right\} dx dx}{\int_0^\infty \exp\left\{-\int_0^x (1+y_{(n)}^2) dx\right\} dx}.$$

This may appear a rather elaborate form of equation (7.50); however it contains the boundary conditions (7.51) in addition, and is in fact quite convenient for numerical work.

Unless $y_{(n+1)} = y_{(n)}$ to the accuracy of the numerical work, $y_{(n+1)}$ is not a solution of equation (7.50), so that the separate members of the sequence $y_{(n)}(x)$ are not solutions of the equation though they do satisfy *all* the boundary conditions. But if the process converges, in a numerical sense that after a finite number of repetitions of the iterative process $y_{(n+1)}$ becomes equal to $y_{(n)}$ to the accuracy to which the numerical work is taken, then to this accuracy such a function $y_{(n+1)}$ is a solution of the equation (7.50).

A process of this kind, when available, is particularly useful in cases in which, using a step-by-step integration, y is very sensitive to y_1 . This sensitiveness is an indication of a kind of instability in the step-by-step process; but this instability does not correspond to any instability in the physical system to which the equation refers, just because in the physical system the behaviour of the system is fixed at both ends of the range. It does not appear either in a treatment such as that of the present section, which approaches the physical situation more closely in that it insists at each stage that the approximate solution is tied at both ends, however it behaves intermediately.

Another method, which also works through a sequence of approximate

solutions satisfying the boundary conditions at both ends of the range, is an application of the 'relaxation' process of Southwell (see § 8.5 and end of § 8.6); and another process is considered in § 8.6.†

7.62. Linear equations with two-point boundary conditions

For linear differential equations with two-point boundary conditions there are special methods which are not applicable to non-linear equations. The equations and boundary conditions together can be divided into three main classes:

(i) Homogeneous equations with conditions $y = 0$ (or more generally homogeneous boundary conditions) at both ends of the range, and the further condition that the solution should not be identically zero.

(ii) Homogeneous equations with the condition y given and non-zero (or more generally an inhomogeneous boundary condition) at one or both ends of the range.

(iii) Inhomogeneous equations.

The first of these classes will be considered in the next section; the present section is concerned with the second and third, the treatments of which are similar.

Consider first the inhomogeneous equation

$$y'' + f(x)y = g(x) \quad (7.53)$$

with $y(a)$ and $y(b)$ given; either or both may be zero without affecting the argument. If Y is any solution of equation (7.53) satisfying the condition $Y = y(a)$ at $x = a$, and z is a solution of the corresponding homogeneous equation

$$z'' + f(x)z = 0 \quad (7.54)$$

satisfying the condition $z = 0$ at $x = a$, then, for any constant α ,

$$y = Y + \alpha z \quad (7.55)$$

is the solution of (7.53) satisfying $y = y(a)$ at $x = a$. In principle, the required solution can be determined by evaluating Y and z by integration of the respective equations and then forming the linear combination (7.55) so as to make $y = y(b)$ at $x = b$; and this may be a practicable procedure. However, if over the range $x = a$ to $x = b$ the solution Y becomes large compared with the solution y to be determined, this procedure may result in the unsatisfactory situation that y is evaluated as the small difference of two relatively large numbers; for example Y

† For a much fuller treatment of numerical methods for differential equations with two-point boundary conditions see L. Fox, *The Numerical Solution of Two-point Boundary Problems in Ordinary Differential Equations* (Clarendon Press, 1957).

and αz might be of the order of 10^4 and y of the order 10^{-4} , and then eight correct significant figures would be needed in Y and z to give one figure in y . Such a situation will usually occur if $b = \infty$ and the complementary function increases without limit as $x \rightarrow \infty$, for example if it behaves asymptotically like x^k ($k > 0$) or e^{kx} ($k > 0$).

In such cases it may be practicable to obtain a solution satisfying the required boundary conditions by integrating outwards from $x = a$ and inwards from $x = b$, and matching the solutions at two intermediate values of x , say $x = X_1$ and $x = X_2 > X_1$ (or perhaps at one intermediate value).

Let Y_{out} be any solution of (7.53) satisfying the condition $Y = y(a)$ at $x = a$, and z_{out} a solution of (7.54) with $z(a) = 0$; then, as before, $Y_{\text{out}} + \alpha z_{\text{out}}$ is a solution of (7.53) satisfying the condition at $x = a$. Also let Y_{in} be any solution of (7.53) satisfying the condition $Y = y(b)$ at $x = b$, and z_{in} a solution of (7.54) with $z(b) = 0$, then $Y_{\text{in}} + \beta z_{\text{in}}$ is a solution of (7.53) satisfying the condition at $x = b$. The outward integration is taken to $x = X_2$ and the inward integration to $x = X_1$, so that they overlap over the range $X_1 \leq x \leq X_2$. To match the results of the inward and outward integrations we require

$$Y_{\text{out}}(X_1) + \alpha z_{\text{out}}(X_1) = Y_{\text{in}}(X_1) + \beta z_{\text{in}}(X_1) \quad (7.56)$$

$$\text{and} \quad Y_{\text{out}}(X_2) + \alpha z_{\text{out}}(X_2) = Y_{\text{in}}(X_2) + \beta z_{\text{in}}(X_2). \quad (7.57)$$

These equations can be solved for α and β , and the solution y constructed from these values, but it may be more convenient to proceed as follows. From (7.56) and (7.57)

$$\frac{Y_{\text{out}}(X_2) - Y_{\text{in}}(X_2) + \alpha z_{\text{out}}(X_2)}{Y_{\text{out}}(X_1) - Y_{\text{in}}(X_1) + \alpha z_{\text{out}}(X_1)} = \frac{z_{\text{in}}(X_2)}{z_{\text{in}}(X_1)},$$

which can be solved for α . Then for $x \leq X_2$, y can be evaluated from

$$y(x) = Y_{\text{out}}(x) + \alpha z_{\text{out}}(x), \quad (7.58)$$

and in particular

$$y(X_1) = Y_{\text{out}}(X_1) + \alpha z_{\text{out}}(X_1).$$

Also for $x \geq X_1$, $y(x) - Y_{\text{in}}(x) = \beta z_{\text{in}}(x)$, so

$$\frac{y(x) - Y_{\text{in}}(x)}{y(X_1) - Y_{\text{in}}(X_1)} = \frac{z_{\text{in}}(x)}{z_{\text{in}}(X_1)}, \quad (7.59)$$

from which y can be constructed for $x > X_1$. The agreement between the values of y , for $X_1 < x < X_2$, calculated from (7.58) and from (7.59) provides a good check that the matching of the results of the inward and outward integration has been carried out correctly.

An alternative matching procedure is to match both the values of y and of y' for the inward and outward integrations at a single intermediate value of x , say $x = X_1$. Then equation (7.57) is replaced by

$$Y'_{\text{out}}(X_1) + \alpha z'_{\text{out}}(X_1) = Y'_{\text{in}}(X_1) + \beta z'_{\text{in}}(X_1),$$

and the subsequent argument is similar. This procedure has two practical disadvantages compared with the above procedure for matching y only, but at two values of x , namely it requires a process of numerical differentiation (§ 6.7) to give the values of the derivatives, and it does not give the check provided by the overlap region $X_1 < x < X_2$ when the solutions are matched at two values of x . However, it may be useful if the Riccati transformation is used for the calculation of complementary function z , since no differentiation is then required to obtain z' , and an alternative check of the matching procedure could be devised.

For a homogeneous equation with $y = y(a) \neq 0$ at $x = a$, or

$$y = y(b) \neq 0 \quad \text{at } x = b,$$

or both, the procedure is very similar, Y_{out} being now a solution of $Y' + f(x)Y = 0$ with $y = y(a)$ at $x = a$, and Y_{in} a solution with $y = y(b)$ at $x = b$; if one or other of $y(a)$ and $y(b)$ is zero, the corresponding function Y is omitted from the argument.

An alternative procedure, based on a matrix treatment of the finite-difference form of the differential equation and boundary conditions, is considered in § 8.6.

7.63. Factorization method

$$\text{For the equation} \quad y'' - k^2 y = g(x) \quad (7.60)$$

with $g(x) \rightarrow 0$ as $x \rightarrow \infty$ and boundary conditions $y \rightarrow 0$ as $x \rightarrow \infty$ and any linear boundary condition at $x = a$, a process based on the factorization of the operator on the left-hand side may be convenient. This factorization gives

$$\left(\frac{d}{dx} - k\right)\left(\frac{d}{dx} + k\right)y = g(x)$$

$$\text{so that if we write} \quad \left(\frac{d}{dx} + k\right)y = v, \quad (7.61)$$

$$\text{then} \quad \left(\frac{d}{dx} - k\right)v = g(x). \quad (7.62)$$

From equation (7.61) it follows that $v \rightarrow 0$ as $x \rightarrow \infty$, so equation (7.62) can be integrated inwards from a known condition for large x , and this integration is stable since the complementary function is e^{+kx} and decreases

as x decreases. This inward integration can be started from the greatest value of x at which $g(x)$ ceases to be negligible to the accuracy required in the calculation. Its value at $x = a$ gives $v(a) = y'(a) + ky(a)$ for the solution satisfying the boundary condition at infinity, and from this value of $v(a)$ and the boundary condition at $x = a$, the value of $y(a)$ can be derived, and forms the boundary condition for the outward integration of equation (7.61); this integration is also stable, since the complementary function is now e^{-kx} , and again decreases in the direction in which the integration is being taken.

This process of treating a second-order equation with two-point boundary conditions by carrying out two integrations, each of a *first-order* equation, over the *whole* range between the values of x , and satisfying one boundary condition at each integration, can be extended† to equations with variable coefficients and inhomogeneous boundary conditions at both ends of a finite range of x .

7.64. Characteristic value problems

In the equation considered in § 7.6, the parameter which is available for adjustment in order that the solution should fit the two-point boundary conditions was y_1 , and in § 7.62 it was the multiple of a complementary function which had to be added to a particular integral. In either case, adjustment of the parameter is equivalent to adjustment of an initial condition in a step-by-step integration.

For a homogeneous linear second-order equation with homogeneous boundary conditions, the adjustable parameter is a constant in the equation itself, for example the constant λ in the equation

$$y'' + [\lambda + f(x)]y = 0 \quad (7.63)$$

with boundary conditions

$$y(a) = y(b) = 0, \quad (7.64)$$

and the further condition that the solution should not be identically zero. For such an equation and boundary conditions there may be no solution (other than $y = 0$) unless λ has one of a set of discrete values, which may be finite or infinite in number. These values of λ are called the 'characteristic values' of the differential equation with these boundary conditions, and the corresponding solutions y are called 'characteristic functions'. Solution of the equation involves determination of one or more of the characteristic values of λ as well as the corresponding characteristic functions; in some cases determination of the characteristic

† See E. C. Ridley, *Proc. Camb. Phil. Soc.* 53 (1957), 442.

values may be more important than determination of the solutions of equation (7.63) themselves.

The determination of characteristic values and functions can be carried out by evaluating trial solutions with trial values of λ . One process is similar to that considered in § 7.62 for an inhomogeneous equation with two-point boundary conditions, by carrying out an outward integration from $x = a$ and an inward integration from $x = b$, and matching them at some intermediate radius, say $x = X$. The results of these two integrations will be written y_{out} and y_{in} .

Since equation (7.63) and the corresponding boundary conditions are homogeneous, it follows that if y is a solution, then so is Ay for any constant value of A . Hence the condition for matching the results of the outward and inward integrations at $x = X$ now consists not of two independent relations $y_{\text{in}}(X) = y_{\text{out}}(X)$ and $y'_{\text{in}}(X) = y'_{\text{out}}(X)$ but the single relation

$$[y'(X)/y(X)]_{\text{in}} = [y'(X)/y(X)]_{\text{out}},$$

expressing the property that if the arbitrary multiplying constants in the inward and outward integrations are so chosen that $y_{\text{in}}(X) = y_{\text{out}}(X)$, then also $y'_{\text{in}}(X)$ must be equal to $y'_{\text{out}}(X)$. If each integration is carried one or two intervals beyond $x = X$, then values of $y'(X)$ can be evaluated from the central-difference formula (6.54) for numerical differentiation; enough figures must be kept in δy to enable these determinations of $y'(X)$ to be carried out to the accuracy required by the rest of the work. Graphs of $y'(X)/y(X)$ against λ for the outward and inward integrations, plotted on the same piece of paper, will assist the choice of successive trial values of λ .

The degree of mismatch between the results of inward and outward integration for a trial value of λ , as measured by the difference between the values of $y'(X)/y(X)$ for the two integrations, may be used directly to estimate an improved trial value.† Let Δy be the difference between two solutions of equation (7.63), both satisfying the condition $y(a) = 0$, with values of λ differing by $\Delta\lambda$, this difference being taken between the values of the two solutions at the same value of x . Then $(y + \Delta y)$ satisfies the equation

$$(y + \Delta y)'' + [(\lambda + \Delta\lambda) + f(x)](y + \Delta y) = 0,$$

so that Δy satisfies the equation

$$\Delta y'' + [(\lambda + \Delta\lambda) + f(x)]\Delta y + (\Delta\lambda)y = 0$$

exactly, or to first order

$$\Delta y'' + [\lambda + f(x)]\Delta y + (\Delta\lambda)y = 0. \quad (7.65)$$

† See E. C. Ridley, *Proc. Camb. Phil. Soc.* **51** (1955), 702.

Now consider the result of integrating equations (7.63) and (7.65) outwards from conditions $y(a) = \Delta y(a) = 0$. By multiplying equation (7.65) by y and (7.63) by $-\Delta y$ and adding, it follows that

$$(y\Delta y'' - y''\Delta y)_{\text{out}} = -(\Delta\lambda)y_{\text{out}}^2. \quad (7.66)$$

The left-hand side of (7.66) is

$$\frac{d}{dx}(y\Delta y' - y'\Delta y)_{\text{out}},$$

and

$$y\Delta y' - y'\Delta y = y^2\Delta(y'/y);$$

also $y(a) = \Delta y(a) = 0$; hence integration between limits $x = a$ and $x = X$ gives

$$\Delta[y'(X)/y(X)]_{\text{out}} = -(\Delta\lambda)\left[\int_a^X y_{\text{out}}^2 dx\right]/[y_{\text{out}}(X)]^2. \quad (7.67)$$

Similarly for the inward integration of equations (7.63) and (7.65) from conditions $y(b) = \Delta y(b) = 0$,

$$\Delta[y'(X)/y(X)]_{\text{in}} = +(\Delta\lambda)\left[\int_X^b y_{\text{in}}^2 dx\right]/[y_{\text{in}}(X)]^2. \quad (7.68)$$

Now if for a trial value of λ the values of

$$[y'(X)/y(X)]_{\text{in}} \quad \text{and} \quad [y'(X)/y(X)]_{\text{out}}$$

do not match, the value of $\Delta\lambda$ required to make them match is the value such that

$$[y'(X)/y(X)]_{\text{out}} + \Delta[y'(X)/y(X)]_{\text{out}} = [y'(X)/y(X)]_{\text{in}} + \Delta[y'(X)/y(X)]_{\text{in}},$$

and substitution from (7.67) and (7.68) gives, to first order in $\Delta\lambda$,

$$\left[\frac{\int_a^X y_{\text{out}}^2 dx}{y_{\text{out}}^2(X)} + \frac{\int_X^b y_{\text{in}}^2 dx}{y_{\text{in}}^2(X)} \right] \Delta\lambda = [y'(X)/y(X)]_{\text{out}} - [y'(X)/y(X)]_{\text{in}}. \quad (7.69)$$

In using this formula, it must be remembered that it is not exact, but only first-order in $\Delta\lambda$; it is most useful for improving an approximation to λ which is already fairly good. In some cases it may be practicable to take $X = b$ in formula (7.69); that is, to dispense with the inward integration.

Another way of matching an inward and an outward integration is to match the ratio of the values of y at two values of X , that is to determine λ so as to make $[y(X_2)/y(X_1)]_{\text{in}} = [y(X_2)/y(X_1)]_{\text{out}}$. This avoids the need for differentiation to give values of y' , but there is then no simple formula

like (7.69) by means of which the degree of the mismatch with one trial value of λ can be used directly to estimate a better value of λ .†

For many forms of the function $f(x)$ for which solutions of (7.63) are wanted in practice, there is a finite least value of λ ; this can often be determined approximately by using Rayleigh's principle.‡ This states that if z is an approximation to the solution of (7.63) for the lowest value of λ , then

$$\int_a^b z\{z'' + f(x)z\} dx \bigg/ \int_a^b z^2 dx \quad (7.70)$$

differs from λ by a quantity of order $\int_a^b (y-z)^2 dx$, so that a rough estimate of z substituted in (7.70) gives a fair value of λ . This may often give a good first trial value to use in the numerical integration of equation (7.63).

Characteristic value problems may also occur with inhomogeneous equations; for example we may require the solution of the equation

$$y'' + [\lambda + f(x)]y = g(x) \quad (7.71)$$

which satisfies the boundary conditions $y(a) = y(b) = 0$ and the 'normalizing' condition $\int_a^b y^2 dx = 1$. For any value of λ which is *not* a

characteristic value of equation (7.63) with the same function $f(x)$ and the same boundary conditions,§ there is a solution satisfying the boundary conditions, and it is unique, but it does not in general satisfy the normalizing condition. The values of λ for which the solution satisfying the boundary conditions also satisfies the normalizing condition are the characteristic values of equation (7.71) with these conditions. They can be found by using the methods of § 7.61 to find the solution of equation (7.71) satisfying the boundary conditions, for a series of trial values of λ ,

evaluating $\int_a^b y^2 dx$ for each solution, and using each result to estimate a better trial value.

† For another procedure, see § 8.7. See also W. E. Milne, *Journ. of Research of Nat. Bur. of Standards*, **45** (1950), 245.

‡ See G. Temple and W. G. Bickley, *Rayleigh's Principle* (Oxford, 1933).

§ If λ is such a characteristic value, equation (7.71) has no solution satisfying the boundary conditions unless the function $g(x)$ satisfies some special conditions.

VIII

SIMULTANEOUS LINEAR ALGEBRAIC EQUATIONS AND MATRICES

8.1. Direct and indirect methods for simultaneous linear equations

THE necessary and sufficient condition that a system of linear simultaneous equations should have a solution, and that the solution should be unique, is that the determinant of the coefficients should be non-zero. In this chapter, except in § 8.7, it will be assumed that this condition is satisfied, so that we shall only be concerned with the determination of a solution when a unique solution exists.

Any textbook of algebra shows that the solution of such a set of equations can be expressed in terms of ratios of determinants, and one way of evaluating the solution of the equations would be to evaluate these determinants numerically. But though there may be no one best way of evaluating the solution, it can be said with some certainty that the direct evaluation of the determinants and of the expression for the solution in terms of them is *never* the best way, though of course the evaluation of a solution by any other method must come in the end to the same thing as the evaluation of the solution in terms of determinants.

The general form of a set of such equations will be written

$$\left. \begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \dots &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \dots &= b_2, \quad \text{etc.} \end{aligned} \right\} \quad (8.1)$$

or shortly
$$\sum_j a_{ij}x_j = b_i \quad (i = 1, 2, \dots, n),$$

or in matrix form
$$\mathbf{Ax} = \mathbf{b}.$$

The equation $\sum a_{ij}x_j = b_i$ with any particular value of i will be called the ' i th equation' and n will be used throughout for the number of equations and of unknowns. \mathbf{I} will be written for the unit matrix, that is the matrix of which all the diagonal elements are 1 and all the non-diagonal elements 0.

There is probably no one way for evaluating the solution of such a set of equations which can be said to be the best in all circumstances. The most effective practical method to use depends on various characteristics of the equations and of the solutions required and on the experience of the individual who has the task of carrying out the numerical work.

The character of the equations is concerned with such matters as whether the coefficients are all small integers or not, whether many are zero, and whether those which are non-zero are arranged in some systematic way, whether they are all exactly known or are subject to uncertainty as a consequence of being either experimental measures or results of other calculations subject to rounding errors, and whether the diagonal coefficients are large compared with the non-diagonal coefficients or not.

Relevant characteristics of the solutions required are whether a solution is wanted for one set of values of the b_j 's only or for many such sets, and whether the characteristic values of the matrix A are required as well as the solution of the equation.

There are two main kinds of method, sometimes called 'direct' and 'indirect'.

'Direct' methods are those in which one application of the computing procedure leads to the solution, to an accuracy depending on the nominal accuracy of the calculations. The evaluation of the expression for the solution in terms of determinants, and the method of elimination, are examples.

'Indirect' methods are those in which the solution is approached by successive approximation, by a number of repetitions of the same computing procedure. For hand calculations an advantage, when such a method is applicable, is that in the early stages only a limited number of figures need be kept, and the accuracy can be increased as the solution is approached.

If an approximation $x_1 = \xi_1$, $x_2 = \xi_2, \dots$, in general $x_i = \xi_i$, to the solution of the equations has been obtained, the quantities

$$R_i = \sum_j a_{ij} \xi_j - b_i \quad (8.2)$$

are called the 'residuals' of the various equations; they measure the extent to which the approximate solution $x_i = \xi_i$ fails to satisfy the equations. The corrections $\Delta \xi_j = (x_j - \xi_j)$ to the approximate solution satisfy the equations

$$\sum_j a_{ij} \Delta \xi_j = c_i = -R_i \quad (8.3)$$

with the same coefficients as the original equations. In various methods of evaluating a solution it is possible to proceed by a process of approximation, obtaining first an approximate solution, calculating the residuals for this approximate solution, then solving (8.3) for corrections to the approximate solution, and perhaps repeating this process.

8.11. Matrices

If, as assumed, the determinant of the coefficients is non-zero then the matrix \mathbf{A} of the coefficients has an inverse \mathbf{A}^{-1} , and the solution is formally

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}. \quad (8.4)$$

If many solutions for different sets of values b_i are wanted, then it may be worth while to evaluate \mathbf{A}^{-1} first and then to evaluate solutions from (8.4); one process for the inversion of a matrix is considered in § 8.3 and another in § 8.41. In general the calculation of \mathbf{A}^{-1} will be affected by rounding errors, so the matrix used as \mathbf{A}^{-1} in evaluating a solution from (8.4) will not be exactly the inverse of the matrix \mathbf{A} of the coefficients of the original equations. The results of evaluating (8.4) with this approximate \mathbf{A}^{-1} may therefore have to be treated as an approximation ξ to the solution and improved as explained at the end of the previous section.

A matrix in which all elements below the diagonal are zero is called an 'upper triangular' matrix and one in which all the elements above the diagonal are zero is called 'lower triangular'. The determinant of a triangular matrix is the product of its diagonal elements.

8.12. Ill-conditioned equations

If the determinant D of the coefficients is expanded in terms of the elements of any one row or column, for example

$$D = \sum_k A_{jk} a_{jk} \quad (8.5)$$

(the sum being over k only, j being fixed), it may happen that D is small compared with some of the individual terms in this sum. Then the value of D , and so the solution of the equations, is very sensitive to small changes in the values of the coefficients, and an approximate solution obtained by a numerical method which is subject to rounding errors is likely to be very sensitive to these errors. If, for example, D is of the order of unity but some of the individual terms in (8.5) are of the order of 2000, then a change of 0.1 per cent. in the coefficient a_{jk} in one of these large terms may change the value of D from $+1$ to -1 , and the solutions in the two cases, and for intermediate values of this coefficient a_{jk} , may be entirely different.

A very elementary example is provided by the equations

$$x + 2y = 4, \quad 1000x + 2001y = 4003$$

for which $D = 2001 - 2000 = 1$, and the solution is

$$x = -2, \quad y = 3.$$

If the coefficient of y in the second of these equations is changed by -0.1 per cent. to 1999, the solution becomes

$$x = 10, \quad y = -3$$

and if it is changed by $+0.1$ per cent. to 2003, the solution becomes

$$x = 2, \quad y = 1.$$

If the coefficients in these equations are known *exactly*, then the solution can be determined to any accuracy required. But if the coefficients are subject to some uncertainty, either through being derived from observations which can only be made to a finite degree of accuracy, or through being themselves results of other calculations which may be affected by rounding errors, then clearly not even the sign and first significant figure of the solution can be determined unless the uncertainties in the coefficients are less than 0.1 per cent. A set of equations for which D is small compared with some of the individual terms in the sum (8.5) is called 'ill-conditioned'.

It is sometimes said that a set of equations is ill-conditioned if the determinant D of the coefficients is small; but this is an inadequate statement because the relevant standard of smallness is unspecified. Consider a set of thirty equations for which $D = 1$. The relations between the variables expressed by the equations are not altered if each equation is multiplied by 1000; but D then becomes 10^{90} , which is not 'small' in any ordinary use of the word.

'Ill-conditioned', applied to a set of equations, is sometimes used merely as a qualitative term of abuse; but it is capable of being given a quantitative significance. Let $\lambda^{(1)}, \lambda^{(2)}, \dots, \lambda^{(n)}$ be the characteristic values of the matrix of the coefficients a_{jk} , and let $|\lambda^{(m)}|$, $|\lambda^{(M)}|$ be the greatest and least of the quantities $|\lambda^{(j)}|$; then $|\lambda^{(M)}|/|\lambda^{(m)}|$ is a quantitative measure of ill-conditionedness; when this ratio is nearly unity the equations are well-conditioned, when it is large compared with 1 they are ill-conditioned; when $\lambda^{(m)}$ is zero the determinant D of the coefficients is zero and either the equations have no solution or the solution is not unique.

Unfortunately the numerical determination of this measure of ill-conditionedness is as long a process as the evaluation of the solution of the equations, and the same applies to some other measures of condition which have been suggested by Turing.[†] So it is not very useful in practice

[†] A. M. Turing, *Quart. J. Mech. and Applied Math.* **1** (1948), 287.

for giving advance warning that the equations concerned are ill-conditioned. In many cases the intermediate results obtained in the course of the numerical process of solving the equations exhibit characteristic symptoms when the equations are ill-conditioned. These symptoms depend on the particular process and will be mentioned in the course of consideration of the individual processes.

Sometimes inspection of a set of equations will suggest that they are ill-conditioned. Expressed geometrically, any one of the equations is the equation of a hyper-plane in n -dimensional space, and the coefficients in the equation are the components of a vector normal to this hyper-plane. If these normals are all in much the same direction then the hyper-planes are nearly parallel so that their intersections are at very acute angles, and the common point of them all, which represents the solution of the equations, is not well determined. For example, it is clear on inspection that for the equations (constructed by T. S. Wilson and quoted by J. Morris)[†]

$$\left. \begin{aligned} 5x_1 + 7x_2 + 6x_3 + 5x_4 &= 23 \\ 7x_1 + 10x_2 + 8x_3 + 7x_4 &= 32 \\ 6x_1 + 8x_2 + 10x_3 + 9x_4 &= 33 \\ 5x_1 + 7x_2 + 9x_3 + 10x_4 &= 31 \end{aligned} \right\} \quad (8.6)$$

the normals to the hyper-planes make only small angles with each other, so that this set of equations is ill-conditioned. But it is only occasionally that an ill-conditioned set of equations can be recognized as such by inspection.

A characteristic feature of ill-conditioned equations is that a set of values for the unknowns which differs considerably from the solution of the equations may, nevertheless, give small residuals for all the equations. For example, for the equations (8.6) the residuals for certain sets of values of x_1, x_2, x_3, x_4 are as follows:

| x_1 | x_2 | x_3 | x_4 | |
|-------|-------|-------|-------|----------------------------------|
| +14.6 | -7.2 | -2.5 | +3.1 | $R_1 = -R_2 = -R_3 = R_4 = 0.1$ |
| +2.36 | +0.18 | +0.65 | +1.21 | $R_1 = -R_2 = -R_3 = R_4 = 0.01$ |

whereas the exact solution is $x_1 = x_2 = x_3 = x_4 = 1$. Thus in this case values of the residuals which are less than $1/2000$ of the values of the b 's in the equations still do not guarantee the accuracy even of the first figure in the x 's. This is an extreme case, but it illustrates the need for caution in taking the smallness of residuals as a guide to the accuracy of the solution when the equations are ill-conditioned.

[†] J. Morris, *Phil. Mag.* (7) **37** (1946), 106.

8.13. Normal equations

For a set of values of (x_1, x_2, \dots, x_n) , *not* necessarily a solution of the equations, let $2S$ be written for the sum of the squares of the residuals (8.2), that is,

$$S = \frac{1}{2} \sum_i R_i^2.$$

Then S is a quadratic function of (x_1, x_2, \dots, x_n) which is zero for those values of (x_1, x_2, \dots, x_n) which form the solution of the equations, and is positive for all other sets of values. Hence the determination of the solution of these equations is equivalent to finding the set of values of (x_1, x_2, \dots, x_n) which make S a minimum. This set is given by

$$\frac{\partial S}{\partial x_j} = 0 \quad (\text{all } j)$$

$$\text{or} \quad \sum_i R_i \frac{\partial R_i}{\partial x_j} = 0. \quad (8.7)$$

$$\text{Now} \quad \partial R_i / \partial x_j = a_{ij} = (\tilde{a})_{ji},$$

where $\tilde{\mathbf{A}}$ is the transpose of the matrix \mathbf{A} . Hence the set of equations (8.7) is

$$\sum_i \tilde{a}_{ji} R_i = 0,$$

$$\text{or} \quad \sum_{ik} [(\tilde{a}_{ji} a_{ik}) x_k - \tilde{a}_{ji} b_i] = 0; \quad (8.8)$$

$$\text{or in matrix form} \quad (\tilde{\mathbf{A}}\mathbf{A})\mathbf{x} - \tilde{\mathbf{A}}\mathbf{b} = 0.$$

The set of equations (8.8) is sometimes called the set of 'normal equations' corresponding to the original equations (8.1). They are derived from the positive definite quadratic form S , and the matrix $(\tilde{\mathbf{A}}\mathbf{A})$ is necessarily symmetrical; these features are advantageous in some methods of carrying out the solution numerically.

On the other hand, the ratio $|\lambda^{(M)}/\lambda^{(m)}|$ for the normal equations is greater than that for the original equations,† so that the normal equations are less well-conditioned than the original equations, and when the original equations are at all severely ill-conditioned, the normal equations are very much worse. Hence, methods which do not depend on the matrix of the coefficients being symmetrical, or on the equations being derived from a positive definite quadratic form, are often preferable.

8.2. Elimination

A straightforward direct method is one based on successive elimination of the variables. This is a development of the process of elementary algebra with attention given to three points, namely systematic arrange-

† See Olga Taussky, *M.T.A.C.* 4 (1950), 111.

ment of the work, provision of a current check, and control of rounding errors. The points are all important and become more important the greater the number of equations and unknowns.

In elementary algebra emphasis is properly given to the importance of verifying that an alleged solution does actually satisfy the equations. This final check should always be carried out. But in the evaluation of the solution of a set of more than three equations it is hardly sufficient since if it fails it gives no indication of the location of the mistake, and the whole work has to be repeated with an appreciable probability of repeating the same mistake. A *current* check is required, both to help in locating a mistake and to prevent much further work being based on erroneous intermediate results.

To eliminate x_k between the i th and j th equations, in which the coefficients of x_k are a_{ik} and a_{jk} respectively, we have to multiply the equations by α_j and α_i respectively, and add, choosing α_i and α_j so that

$$\alpha_j a_{ik} + \alpha_i a_{jk} = 0.$$

Formally, there is an infinite number of ways of choosing the multipliers α_i and α_j ; but in practice there are only two ways which are generally useful. One is to take one of the multipliers as unity, that is to say to take

$$\left. \begin{aligned} \alpha_i &= 1, & \alpha_j &= -a_{jk}/a_{ik} \\ \alpha_j &= 1, & \alpha_i &= -a_{ik}/a_{jk} \end{aligned} \right\} \quad (8.9)$$

or

the other is to take

$$\alpha_j = \pm a_{jk}, \quad \alpha_i = \mp a_{ik}. \quad (8.10)$$

The division involved in the choice represented by (8.9) will usually involve rounding errors in each elimination. Since the results of eliminating one variable are used later in the elimination of other variables, it is important to keep these rounding errors under control, and this is best done as follows. One of the multipliers is taken as unity; we choose that one of the alternatives (8.9) which makes the modulus of the other multiplier less than unity. Then at each stage of the elimination the rounding errors from previous stages are always multiplied by numbers of modulus less than unity.

This is a general method. If, however, the coefficients are all fairly small integers, it is possible to carry out the elimination without introducing any rounding errors, by using the choice (8.10) of the multipliers, which avoids any division. If this can be done without the coefficients becoming inconveniently large it is probably the preferable choice. It may be possible to keep the coefficients from becoming large by making suitable linear combinations of the equations, with integral coefficients,

in the course of the elimination process; since no rounding off is involved there is no loss of numerical accuracy in such a procedure. As a simple example, with the equations

$$(1) \ 23x+31y = b_1, \quad (2) \ 44x+65y = b_2$$

one should *not* try to eliminate x by multiplying the first equation by 44 and the second by -23 and adding; the first step should be to form the linear combination $2 \times (1) - (2)$:

$$(3) = 2 \times (1) - (2), \quad 2x - 3y = 2b_1 - b_2,$$

before continuing the elimination.

8.21. General elimination process

To eliminate the variable x_k from a set of equations, specified by different values of i , we may take one equation, say the j th, and for each value of i form

$$(i\text{th equation}) + (\alpha_i) \times (j\text{th equation}), \quad (8.11)$$

where α_i is given by the second of formulae (8.9). If the numerical work is suitably arranged, it is not necessary to write out the equations in full at each stage. It is enough to write down, for each equation, the coefficients and the constant term in appropriate columns. A current check can be provided by keeping a record, with each equation, of the sum of the coefficients and the constant term, and forming the linear combination expressed by (8.11) not only for the coefficients but also for this check sum. The value of this check sum for the i th equation will be written s_i ; that is

$$s_i = b_i + \sum_{\bar{u}} a_{i\bar{u}}.$$

Thus for using the j th equation to eliminate x_k from the i th equation, we have the following scheme:

| | Coefficients of | | Constant | Check |
|-------------------------------|----------------------------|----------------------------|----------------------|----------------------|
| | x_k | x_l | term | sum |
| i th equation | a_{ik} | a_{il} | b_i | s_i |
| j th equation | a_{jk} | a_{jl} | b_j | s_j |
| $[\alpha_i = -a_{ik}/a_{jk}]$ | $a_{ik} + \alpha_i a_{jk}$ | $a_{il} + \alpha_i a_{jl}$ | $b_i + \alpha_i b_j$ | $s_i + \alpha_i s_j$ |
| Result | $= 0$ | $= a'_{il}$ | $= b'_i$ | $= s'_i$ |

(8.12)

The check consists in verifying that s'_i calculated as $s_i + \alpha_i s_j$ is in agreement (within the tolerance for rounding errors) with the sum of the other entries in the corresponding line, namely $b'_i + \sum_{\bar{l}} a'_{i\bar{l}}$. The j th equation here is sometimes called the 'pivotal equation', and the coefficient a_{jk} , which is the divisor in the evaluation of the coefficient $\alpha_i = -a_{ik}/a_{jk}$,

is called the ‘pivotal coefficient’ or ‘pivot’ for this elimination; it is the coefficient, in the pivotal equation, of the variable to be eliminated.

Example: To use the first of the three equations:

$$\begin{aligned} 31.74x_1 + 43.61x_2 - 16.94x_3 + 16.94x_4 &= 41.37, \\ 6.86x_1 + 9.81x_2 + 7.68x_3 + 3.96x_4 &= 16.81, \\ 35.85x_1 - 32.92x_2 + 13.81x_3 + 5.94x_4 &= 21.84 \end{aligned}$$

to eliminate x_2 from the second and third.

The pivotal coefficient is the coefficient 43.61 in the first equation, and the multipliers are

$$\alpha_2 = -9.81/43.61 = -0.22495, \quad \alpha_3 = +32.92/43.61 = 0.75487$$

| Line no. and operation | Coefficient of | | | | Constant term | Check sum | Notes |
|------------------------------|----------------|--------|--------|---------|------------------|--------------|------------------------|
| | x_1 | x_2 | x_3 | x_4 | | | |
| (1) | 31.74 | 43.61 | -16.94 | 16.94 | 41.37 | 116.72 | |
| (2) | 6.86 | 9.81 | 7.68 | 3.96 | 16.81 | 45.12 | |
| (3) | 35.85 | -32.92 | 13.81 | 5.94 | 21.84 | 44.52 | |
| (4) = $\alpha_2 \times (1)$ | -7.140 | -9.810 | 3.811 | -3.811 | -9.306 | -26.256 | $\alpha_2 = -0.22495$ |
| (5) = (2) + (4) | -0.280 | 0 | 11.491 | 0.149 | 7.504 | 18.864 | cross sum = 18.864 |
| (6) = $\alpha_3 \times (1)$ | 23.960 | 32.920 | 12.877 | -12.877 | 31.229 | 88.108 | $\alpha_3 = 0.75487$ |
| (7) = (3) + (6) | 59.810 | 0 | 26.687 | -6.937 | 53.069 | 132.628 | cross sum = 132.629 |

Notes: (i) It is convenient to keep a note, on the left-hand side, of the operations carried out to obtain the successive lines of the calculation. This simplifies the location and correction of mistakes should any be made, and is also useful if the calculation has to be repeated with other values of the constant terms b_i .

(ii) In working, it is best to set each α_i on the setting levers or keyboard of a machine, then first multiply by the coefficient a_{jk} of the variable which it is required to eliminate, to check that α_i has been set correctly, then multiply by the other coefficients a_{ji} in the pivotal equation. These products can be written down, as in the above example in lines (4) and (6), or the entries in lines (5) and (7) can be formed and written down directly.

(iii) The results of elimination of x_2 from the third equation in the above example illustrate a kind of mistake which is not detected by the check explained above. The check is satisfied but two entries in line (7) are wrong; in line (6) the coefficients of x_3 and x_4 should read -12.787 and $+12.787$ respectively, and the corresponding entries in line (7) need correcting accordingly. The failure of the check to indicate the presence of the mistakes arises from the fact that the coefficient of x_4 in line (6) is just copied from that of x_3 with a change of sign, and if the latter coefficient is wrong, then so is the former and the effects of the two mistakes cancel in the check. This shows that special care is necessary when two coefficients in the pivotal equation are equal and opposite in sign. It also emphasizes that in the solution of a system of equations the current check on the eliminations alone is not sufficient; this current check should *always* be supplemented by a verification that an alleged solution does satisfy the original equations.

8.22. Evaluation of a solution by elimination

One way of arranging the solution of a system of equations by the elimination process is as follows. Use one equation as pivotal equation to eliminate one variable from all the equations. It is convenient to consider the variables and equations renumbered (if necessary) so that the variable eliminated is x_1 , and the equation used as pivotal equation at this stage is the first. The result will be a single equation containing x_1 :

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \dots + a_{1n}x_n = b_1 \quad (8.13)$$

and a system of $(n-1)$ equations not involving x_1 , of which a typical one is

$$a'_{i2}x_2 + a'_{i3}x_3 + \dots + a'_{in}x_n = b'_i, \quad (8.14)$$

where a'_{ij} and b'_i are given by (8.12) with $j = 1$. Then another variable, which can similarly be taken as x_2 , is eliminated from equations (8.14), and so on.

If the equation used as the pivotal equation to eliminate x_1 is chosen so that $|a_{11}| \geq |a_{i1}|$ ($i > 1$), then none of the multipliers $\alpha_i = -(a_{i1}/a_{11})$ in this elimination is greater than 1 in magnitude, a condition which we have already seen to be desirable to keep control of rounding errors. Similarly, the equation used as the pivotal equation to eliminate x_2 should be chosen so that $|a'_{22}| \geq |a'_{i2}|$ ($i > 2$), and so on.

Once an equation has been used as a pivotal equation in the elimination of one variable in this process, it is left unaltered in the further stages of the elimination process. Then this process leads finally to a system of equations of which the first is (8.13), the second is that member (say with $j = 2$) of the set (8.14) which is used as the pivotal equation for the elimination of x_2 from this set of equations, and so on. This system is

$$\left. \begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + a_{14}x_4 + \dots &= b_1 \\ a'_{22}x_2 + a'_{23}x_3 + a'_{24}x_4 + \dots &= b'_2 \\ a''_{33}x_3 + a''_{34}x_4 + \dots &= b''_3, \text{ etc.} \end{aligned} \right\}, \quad (8.15)$$

of which the m th equation contains $n-m+1$ of the variables, and the coefficients form an upper triangular matrix.

The last of this set of equations involves x_n only, the last but one involves x_n and x_{n-1} only, the last but two involves x_n , x_{n-1} , and x_{n-2} only, and so on. This whole system of equations can therefore be solved by starting from the *last* and working backwards so as to determine the values of x_n , x_{n-1} , x_{n-2} , ... in this order, using at each stage the values of the x_k 's previously obtained. This process is known as 'back substitution'. In it, the value of each unknown is determined from that

equation which was used as pivotal equation to eliminate this unknown in the elimination process.

In the calculation the usual symptom of an ill-conditioned set of equations is that the coefficient of x_n in the last equation of the set (8.15) is small compared with the values of the coefficients in the equation used for the elimination of x_{n-1} ; this is illustrated in the following example. Another symptom (of which this is a particular case) which may appear is that the elimination of one of the unknowns between two equations, say equations (A) and (B), by forming the linear combination $(A)+\alpha(B)$ (with $|\alpha| \leq 1$), results in an equation in which *all* the coefficients have values which are small compared with their values in equation (A).

Example:

To solve the equations

$-23x_1+11x_2+x_3 = 0, \quad 11x_1-3x_2-2x_3 = 3, \quad x_1-2x_2+x_3 = -2,$

| Line no. and operation | Coefficients of | | | b | Check sum | Notes |
|------------------------------|-----------------|--------|--------|---------|--------------|---|
| | x_1 | x_2 | x_3 | | | |
| (1) | -23 | 11 | 1 | 0 | -11 | |
| (2) | 11 | -3 | -2 | 3 | 9 | |
| (3) | 1 | -2 | 1 | -2 | -2 | |
| (4) = $\alpha_1 \times (1)$ | -11 | 5.261 | 0.478 | 0 | -5.261 | $\alpha_1 = 11/23 = 0.47826$ |
| (5) = (2)+(4) | | 2.261 | -1.522 | 3 | 3.739 | |
| | | | | | 3.739 | cross sum |
| (6) = $\alpha_2 \times (1)$ | -1 | 0.478 | 0.043 | 0 | -0.478 | $\alpha_2 = 1/23 = 0.04348$ |
| (7) = (3)+(6) | | -1.522 | 1.043 | -2 | -2.478 | |
| | | | | | -2.479 | cross sum |
| (8) = $\alpha_3 \times (5)$ | | 1.522 | -1.025 | 2.020 | 2.517 | $\alpha_3 = \frac{1.522}{2.261} = 0.6732$ |
| (9) = (7)+(8) | | | 0.018 | 0.020 | 0.038 | |
| | | | | | 0.038 | cross sum |
| (10) = x_3 | | | 1 | 1.111 | 2.111 | |
| (11) = $1.522x_3$ | | | 1.522 | 1.691 | 3.213 | |
| (12) = (5)+(11) | | 2.261 | | 4.691 | 6.952 | |
| (13) = x_3 | | 1 | | 2.075 | 3.075 | cross sum checks |
| (14) = $-11 \times (13)$ | | -11 | | -22.825 | -33.825 | |
| (15) = (1)+(14) | | | | | | |
| -(10) | -23 | | | -23.936 | -46.936 | |
| (16) = x_1 | 1 | | | 1.041 | 2.041 | |

| Final check | | Residuals |
|---------------|----------------------|-----------|
| $x_1 = 1.041$ | $-23x_1+11x_2+x_3 =$ | -0.007 |
| $x_2 = 2.075$ | $11x_1-3x_2-2x_3 =$ | +0.004 |
| $x_3 = 1.111$ | $x_1-2x_2+x_3 =$ | +0.002 |

Notes: (i) The 'pivots' are distinguished in this example by being printed in heavy type (in manuscript an underline or 'box' could be used), the coefficient of the last remaining variable being counted as a 'pivot' for this purpose. The largest coefficient is chosen as the first pivot, and the result of eliminating x_1 is given by lines (5) and (7). In these equations the coefficient of x_2 in line (5) is the greatest, and this is taken as the next pivot.

(ii) In line (7) there is a difference of a unit in the last figure between the number derived from the other relevant entries in the same *column* and that derived from the cross sum along the *row*. This is an effect of rounding errors, and occasional small discrepancies in the check, such as this, must be expected. In the further calculation, the value derived from the cross sum should be used in such cases.

(iii) If the coefficients are known to have exact integral values, then it is significant to keep any number of figures in the calculation. In this example three decimals have been kept. Any required accuracy could be attained by keeping enough figures, though before the solution is carried out it may be difficult to judge how many are needed to give a specified accuracy in the solution. In the present case it might be expected that the three decimals kept should be enough to give the solution to 1 per cent., allowing for rounding errors; but as will be seen (under (iv) below), the 'solution' in this case is not even accurate to two figures.

If the coefficients were only known to two decimals there would be no significance in keeping more than three. If, for example, the coefficient a_{22} were only known to lie in the range -3.00 ± 0.01 , then the entry 2.261 in line (5) might stand for any number in the range 2.251 to 2.271, and the only purpose of keeping even the third decimal is to avoid rounding errors accumulating in the second decimal.

(iv) By taking another decimal in x_1 only ($x_1 = 1.0407$) these residuals can be reduced to -0.001 , $+0.001$, $+0.002$ to three decimals, so that at first sight it would appear that this solution is correct to at least two decimals. It is not, however: the correct solution is $x_1 = 1$, $x_2 = 2$, $x_3 = 1$, and the solution is not even correct to one decimal. This is a consequence of the ill-conditioned character of the equations. There is a warning of this character of the equations at line (9) of the working, at which the coefficient of x_3 is very small compared with that in equation (7).

(v) If the coefficients are known to have exact integral values, the value of the last pivotal coefficient can sometimes be improved as follows. The product of the pivots is the determinant of the coefficients of the original equations,† and this must be an integer if the coefficients are integers. In this example the product of the pivots as evaluated is

$$-23 \times 2.261 \times 0.018 = -0.94.$$

The second factor is certainly not in error by more than 1 in 1000, and the extreme possibilities of rounding errors cannot affect the third by more than 10 per cent.; so the value of this product lies between -0.84 and -1.04 . But it must be integral, and must therefore be -1 ; hence the value of the last pivot is

$$-1/(-23 \times 2.261) = 0.0192,$$

the fourth decimal being certainly correct.

This argument must *not* be used unless the coefficients are *known* to be integral; otherwise it might give a false idea of the accuracy of the solution.

† A factor (-1) may be introduced if the order of the variables or of the equations is changed in the elimination process.

(vi) If ξ_1, ξ_2, ξ_3 is an approximate solution, and R_1, R_2, R_3 the residuals obtained by putting $x_j = \xi_j$, then the corrections $(x_j - \xi_j)$ to the approximate solution can be obtained by solving the equations (8.3) in a similar manner.

8.23. Alternative arrangement of the elimination process

In the above arrangement of the elimination process, each pivotal equation is left unchanged in later stages of the process. An alternative procedure is to use each pivotal equation to eliminate an unknown from the pivotal equations previously used, as well as from later equations.

In the above example, for instance, the pivotal equation (5) can be used to eliminate x_2 from equation (1) as well as from equation (3).

This procedure avoids the process of back-substitution, but the elimination process is longer, and the total amount of work involved is about the same.

8.3. Inverse of a matrix by elimination

One way of inverting a matrix \mathbf{A} is to obtain a set of solutions of the equations

$$\sum_j a_{ij} x_j = b_i \quad (8.16)$$

for a general set of values of the b 's. Each step of the elimination process consists of forming a linear combination of the left-hand sides of the various equations (8.16), so the individual x 's which result from the elimination process can be expressed as linear combinations of the left-hand sides. If the same linear combinations of the right-hand sides are formed, the result is that each x_j is expressed as a linear combination of the b_i 's, say

$$x_j = \sum_i c_{ji} b_i \quad (8.17)$$

or in matrix form $\mathbf{x} = \mathbf{C}\mathbf{b}$, where \mathbf{C} is the matrix of the coefficients in (8.17). But if \mathbf{A} is non-singular, there is a unique matrix \mathbf{A}^{-1} such that $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$; hence the matrix \mathbf{C} is just the inverse of \mathbf{A} .

The solution of the equations for general values of the b 's can be carried out by a slight extension of the process of § 8.22, using a separate column for the coefficient of each b_k .

Example: To solve the equations

$$-23x_1 + 11x_2 + x_3 = b_1,$$

$$11x_1 - 3x_2 - 2x_3 = b_2,$$

$$x_1 - 2x_2 + x_3 = b_3$$

for general values of b_1, b_2 , and b_3 . The left-hand sides of these equations are the same as in the previous example; the right-hand sides have general values. In this example the elimination will be carried out by the special process mentioned in § 8.21 as avoiding divisions, and the rounding errors associated with them, in the elimination process.

| Equation no. and operation | Coefficients of | | | Coefficients of | | | Check sum | Notes |
|---------------------------------|-----------------|-------|-------|-----------------|-------|-------|--------------|------------------|
| | x_1 | x_2 | x_3 | b_1 | b_2 | b_3 | | |
| (1) | -23 | 11 | 1 | 1 | 0 | 0 | -10 | |
| (2) | 11 | -3 | -2 | 0 | 1 | 0 | 7 | |
| (3) | 1 | -2 | 1 | 0 | 0 | 1 | 1 | |
| (4) = (1) - (3) | -24 | 13 | 0 | 1 | 0 | -1 | -11 | cross sum checks |
| (5) = (2) + 2 × (3) | 13 | -7 | 0 | 0 | 1 | 2 | 9 | cross sum checks |
| (6) = (4) + 2 × (5) | 2 | -1 | 0 | 1 | 2 | 3 | 7 | cross sum checks |
| (7) = (4) + 12 × (6) = x_1 | 0 | 1 | 0 | 13 | 24 | 35 | 73 | cross sum checks |
| (8) = (6) + (7) | 2 | 0 | 0 | 14 | 26 | 38 | 80 | |
| (9) = $\frac{1}{2}$ (8) = x_1 | 1 | 0 | 0 | 7 | 13 | 19 | 40 | cross sum checks |
| (10) = 2 × (7) + (3) | 1 | 0 | 1 | 26 | 48 | 71 | 147 | |
| (11) = (10) - (9) = x_3 | 0 | 0 | 1 | 19 | 35 | 52 | 107 | cross sum checks |

Hence

$$x_1 = 7b_1 + 13b_2 + 19b_3,$$

$$x_2 = 13b_1 + 24b_2 + 35b_3,$$

$$x_3 = 19b_1 + 35b_2 + 52b_3,$$

and

$$\begin{bmatrix} -23 & 11 & 1 \\ 11 & -3 & -2 \\ 1 & -2 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 7 & 13 & 19 \\ 13 & 24 & 35 \\ 19 & 35 & 52 \end{bmatrix}.$$

Notes: (i) In this example advantage has been taken of the simple numerical values of the coefficients to lighten the numerical work of the elimination process. The particularly simple values of the coefficients of x_3 suggest that this is the unknown to eliminate first.

(ii) In line (6) no elimination is carried out, but a linear combination of the equations is made so as to keep down the magnitudes of the numbers occurring in the calculation.

(iii) By avoiding division and so keeping the work free from rounding errors, the exact solution is obtained without any attention having to be given to the number of figures kept at the various stages of the work. Further, the ill-conditioned nature of the equations (see note (iv) below) gives no difficulty in obtaining a solution. Also the numbers occurring are simple enough in this case for the whole calculation to be done without the aid of a desk machine.

(iv) The large values of the elements of the inverse matrix show why such a poor approximation to the solution, as represented by the 'solution' obtained in § 8.22, gives such small residuals.

If (ξ_1, ξ_2, ξ_3) is an approximation to the solution, and R_1, R_2, R_3 are the residuals obtained on substituting $x_1 = \xi_1, x_2 = \xi_2, x_3 = \xi_3$ into the equations, then the corrections to the approximate solution are

$$(x_1 - \xi_1) = 7R_1 + 13R_2 + 19R_3,$$

$$(x_2 - \xi_2) = 13R_1 + 24R_2 + 35R_3,$$

$$(x_3 - \xi_3) = 19R_1 + 35R_2 + 52R_3,$$

so that if $R_1 = R_2 = R_3 = 0.01$, then $x_3 - \xi_3 = 1.06$; that is, the error in an approximate value of x_3 may be over 100 times the residuals in the equations, although in the equations this unknown only occurs with coefficients 1 and 2.

8.4. Choleski's method

An alternative direct method is one usually ascribed to Choleski,[†] of which there are several variants. It depends on the factorization of the matrix **A** of the coefficients into the product **LU** of two matrices, of which **L** is lower triangular and **U** is upper triangular. In such a factorization, the diagonal elements of either **L** or **U** (but not of both) can be restricted to be unity. Then the system of equations

$$\mathbf{Ax} = \mathbf{LUx} = \mathbf{b}$$

$$\text{can be written} \quad \mathbf{Ly} = \mathbf{b}, \quad (8.18)$$

$$\mathbf{Ux} = \mathbf{y}. \quad (8.19)$$

Written out, equations (8.18) are

$$\left. \begin{aligned} l_{11}y_1 &= b_1 \\ l_{21}y_1 + l_{22}y_2 &= b_2 \\ l_{31}y_1 + l_{32}y_2 + l_{33}y_3 &= b_3 \\ \cdot &\quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \end{aligned} \right\} \quad (8.20)$$

From these equations, y_1, y_2, y_3, \dots can be obtained in succession. Equations (8.19), written out, are

$$\left. \begin{aligned} u_{11}x_1 + u_{12}x_2 + u_{13}x_3 + \dots + u_{1n}x_n &= y_1 \\ u_{22}x_2 + u_{23}x_3 + \dots + u_{2n}x_n &= y_2 \\ \cdot &\quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\ u_{n-1,n-1}x_{n-1} + u_{n-1,n}x_n &= y_{n-1} \\ u_{nn}x_n &= y_n \end{aligned} \right\} \quad (8.21)$$

and, once the values of the y_j 's have been found from (8.20), those of the x_k 's can be found in succession from equations (8.21), starting with the *last* and working backwards.

From equations (8.20) it follows that the y 's are given by

$$\left. \begin{aligned} y_1 &= b_1/l_{11} \\ y_2 &= (b_2 - l_{21}y_1)/l_{22} \\ y_3 &= (b_3 - l_{31}y_1 - l_{32}y_2)/l_{33} \end{aligned} \right\} \quad (8.22)$$

and one way of determining them is to evaluate these formulae in succession, using in each one the values of the y 's determined from previous formulae in the sequence. Wilkes[‡] has given an alternative way of arranging the work, and has shown how it can be developed to give a convenient practical procedure for carrying out the factorization of **A**

[†] See, for example, L. Fox, H. D. Huskey, and J. H. Wilkinson, *Quart. J. Mech. and Applied Math.* **1** (1948), 149; A. M. Turing, *ibid.* 287. For another method of a similar type see P. D. Crout, *Trans. Amer. Inst. Elect. Eng.* **60** (1941), 1235.

[‡] M. V. Wilkes, *Proc. Camb. Phil. Soc.* **52** (1956), 758.

in the form $\mathbf{A} = \mathbf{LU}$. This uses a scheme similar to that of division of one polynomial by another by the method of detached coefficients.

Consider the result of subtracting from the sequence (b_1, b_2, \dots, b_n) such a multiple of the sequence $(l_{11}, l_{21}, \dots, l_{n1})$ that the first term is reduced to zero. This can be laid out as follows:

$$\begin{array}{cccccccc} l_{11} & l_{21} & l_{31} & \dots & b_1 & & b_2 & & b_3 & \dots & (b_1/l_{11} = y_1) \\ & & & & b_1 & & l_{21}y_1 & & l_{31}y_1 & \dots & \\ \hline & & & & 0 & & b_2 - l_{21}y_1 & & b_3 - l_{31}y_1 & \dots & \end{array}$$

the 'quotient' being y_1 , and the 'remainder' sequence being the numbers $b'_j = b_j - l_{j1}y_1$ ($j \geq 2$) which would be obtained on the right-hand side of equations (8.20) by using the first of these equations to eliminate y_1 from the rest. Now 'divide' this remainder sequence, in the same way, by the second column of \mathbf{L} , namely the sequence $(l_{22}, l_{32}, \dots, l_{n2})$:

$$\begin{array}{cccccccc} l_{22} & l_{32} & \dots & b_2 - l_{21}y_1 & & b_3 - l_{31}y_1 & \dots & ([b_2 - l_{21}y_1]/l_{22} = y_2) \\ & & & b_2 - l_{21}y_1 & & l_{32}y_2 & \dots & \\ \hline & & & 0 & & b_3 - l_{31}y_1 - l_{32}y_2 & \dots & \end{array}$$

the 'quotient' being y_2 and the 'remainder' sequence the numbers $b''_j = b_j - l_{j1}y_1 - l_{j2}y_2$ ($j \geq 3$). This 'remainder' is then 'divided' by the third column of \mathbf{L} , and so on.

In this procedure, successive operations are concerned with successive *columns* of \mathbf{L} , whereas in the evaluation of the y 's directly from formulae (8.22), each step is concerned with the evaluation of a formula derived from a *row* of \mathbf{L} . A check is provided by the result of summing by columns. The sum of the equations (8.20) is

$$\left(\sum_j l_{j1}\right)y_1 + \left(\sum_j l_{j2}\right)y_2 + \dots = \sum_j b_j,$$

the sum over j being for the whole set of equations ($y = 1$ to n) or for the first k of them ($j = 1$ to k); the sums over k equations ($k = 2$ to $n-1$) provide current checks. Alternatively, a column-sum check can be kept in the 'division' process. A similar treatment can be applied to the set of equations (8.21), starting with the last column of \mathbf{U} .

Now consider the factorization $\mathbf{A} = \mathbf{LU}$, with \mathbf{U} restricted to have its diagonal elements unity. That is, we require l_{jk} and u_{jk} such that:

$$\begin{bmatrix} l_{11} & & & \\ l_{21} & l_{22} & & \\ l_{31} & l_{32} & l_{33} & \\ \cdot & \cdot & \cdot & \end{bmatrix} \begin{bmatrix} 1 & u_{12} & u_{13} & \cdot & \cdot & \cdot \\ & 1 & u_{23} & \cdot & \cdot & \cdot \\ & & 1 & \cdot & \cdot & \cdot \\ & & & \cdot & \cdot & \cdot \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdot & \cdot & \cdot \\ a_{21} & a_{22} & a_{23} & \cdot & \cdot & \cdot \\ a_{31} & a_{32} & a_{33} & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix} \quad (8.23)$$

The first column of **A** gives

$$l_{j1} = a_{j1} \quad (\text{all } j).$$

The first column of **L** is now known. Then from the second column of **A**

$$\left. \begin{aligned} l_{11} u_{12} &= a_{12} \\ l_{j1} u_{12} + l_{j2} u_{22} &= a_{j2} \quad (j \geq 2) \end{aligned} \right\} \quad (8.24)$$

of which the first gives $u_{12} = a_{12}/l_{11}$, the unknown element in the second column of **U**; since $u_{22} = 1$, the second of equations (8.24) gives

$$l_{j2} = a_{j2} - l_{j1} u_{12} \quad (j \geq 2),$$

the unknown elements in the second column of **L**. The evaluation of these elements can be arranged as a division process:

$$\begin{array}{ccccccc} l_{11} & l_{21} & l_{31} & \dots & a_{12} & a_{22} & a_{32} \dots & (a_{12}/l_{11} = u_{12} \\ & & & & a_{12} & l_{21} a_{12} & l_{31} a_{12} & \dots \\ \hline & & & & 0 & a_{22} - l_{21} a_{12} & a_{32} - l_{31} a_{12} & \dots \\ & & & & & = l_{22} & = l_{32} & \end{array}$$

The first two columns of **L** are now known.

The third column of **A** gives the equations

$$\left. \begin{aligned} l_{11} u_{13} &= a_{13} \\ l_{21} u_{13} + l_{22} u_{23} &= a_{23} \\ l_{j1} u_{13} + l_{j2} u_{23} + l_{j3} u_{33} &= a_{j3} \quad (j \geq 3) \end{aligned} \right\} \quad (8.25)$$

of which the first two give

$$u_{13} = a_{13}/l_{11}$$

and

$$u_{23} = (a_{23} - l_{21} u_{13})/l_{22},$$

the two unknown elements in the third column of **U**; since $u_{33} = 1$, the third of equations (8.25) gives

$$l_{j3} = a_{j3} - l_{j1} u_{13} - l_{j2} u_{23},$$

the unknown elements in the third column of **L**. The calculation can again be arranged like a division, the two 'divisors' being the two columns of **L** previously determined:

$$\begin{array}{ccccccc} l_{11} & l_{21} & l_{31} & \dots & a_{13} & a_{23} & a_{33} \dots & (a_{13}/l_{11} = u_{13} \\ & & & & a_{13} & l_{21} u_{13} & l_{31} u_{13} & \dots \\ \hline & l_{22} & l_{32} & \dots & a_{23} - l_{21} u_{13} & a_{33} - l_{31} u_{13} & \dots & \left(\frac{a_{23} - l_{21} u_{13}}{l_{22}} = u_{23} \right. \\ & & & & a_{23} - l_{21} u_{13} & l_{32} u_{23} & & \\ \hline & & & & & a_{33} - l_{31} u_{13} - l_{32} u_{23} & \dots & \\ & & & & & = l_{33} & \dots & \end{array}$$

The procedure can now be applied to the further columns of \mathbf{A} in succession. In each case the 'divisor' sequences are the successive columns of \mathbf{L} already evaluated, the 'quotients' are the unknown elements in the next column of \mathbf{U} , and the 'remainder' sequence consists of the unknown elements in the next column of \mathbf{L} .

The only divisions of numbers required are divisions by the diagonal elements of \mathbf{L} . Now the determinant of a triangular matrix is the product of its diagonal elements, and so is unity for \mathbf{U} . Hence the determinant of \mathbf{L} is equal to that of \mathbf{A} , which, as stated in § 8.1, is supposed to be non-zero so that the equation $\mathbf{Ax} = \mathbf{b}$ has a unique solution. Hence no diagonal element of \mathbf{L} is zero, and no step of the calculation calls for a division by zero. When the equations are ill-conditioned, however, one diagonal element of \mathbf{L} (or more) will be small compared with the others, and, unless an adequate number of digits have been carried in the intermediate working, the values obtained may be considerably influenced by rounding errors.

Example: To factorize the matrix $\mathbf{A} = \begin{bmatrix} 2 & 0 & -4 & 6 \\ -1 & 1 & 5 & -2 \\ 4 & 0 & -5 & 6 \\ 2 & -1 & -7 & 10 \end{bmatrix}$.

The first column of \mathbf{L} is the same as that of \mathbf{A} . Since in this example $a_{12} = 0$ it follows that $u_{12} = 0$ and the second column of \mathbf{L} consists of the other elements of the second column of \mathbf{A} . Hence

$$\mathbf{L} = \begin{bmatrix} 2 & & & \\ -1 & 1 & & \\ 4 & 0 & \times & \\ 2 & -1 & \times & \times \end{bmatrix}, \quad \mathbf{U} = \begin{bmatrix} 1 & 0 & \times & \times \\ & 1 & \times & \times \\ & & 1 & \times \\ & & & 1 \end{bmatrix},$$

the \times 's standing for elements still unknown, and the blanks for elements known to be zero.

For the third columns of \mathbf{U} and \mathbf{L} , the process of 'division' of the third column of \mathbf{A} by the first and second columns of \mathbf{L} gives

$$\left. \begin{array}{r} 2 \quad -1 \quad 4 \quad 2) \quad -4 \quad 5 \quad -5 \quad -7 \quad (-2) \\ \quad \quad \quad -2 \quad +2 \quad -8 \quad -4 \\ \hline 1 \quad 0 \quad -1) \quad 3 \quad 3 \quad -3 \quad (3) \\ \quad \quad \quad 3 \quad 0 \quad -3 \\ \hline \quad \quad \quad 3 \quad 0 \end{array} \right\} \text{third column of } \mathbf{U}$$

third column of \mathbf{L}

so that

$$\mathbf{L} = \begin{bmatrix} 2 & & & \\ -1 & 1 & & \\ 4 & 0 & 3 & \\ 2 & -1 & 0 & \times \end{bmatrix}, \quad \mathbf{U} = \begin{bmatrix} 1 & 0 & -2 & \times \\ & 1 & 3 & \times \\ & & 1 & \times \\ & & & 1 \end{bmatrix},$$

\times 's indicating elements still unknown.

Then a similar treatment applied to the fourth column of \mathbf{A} gives

$$\left. \begin{array}{rrrrrrrr} 2 & -1 & 4 & 2) & 6 & -2 & 6 & 10 & (3 \\ & & & & 6 & -3 & 12 & 6 & \\ 1 & 0 & -1) & & +1 & -6 & 4 & (1 \\ & & & & 1 & 0 & -1 & & \\ & & & 3 & 0) & -6 & 5 & (-2) \\ & & & & & -6 & 0 & & \\ & & & & & & 5 & = l_{44} \end{array} \right\} \text{fourth column of } \mathbf{U}$$

so that finally

$$\mathbf{L} = \begin{bmatrix} 2 & & & \\ -1 & 1 & & \\ 4 & 0 & 3 & \\ 2 & -1 & 0 & 5 \end{bmatrix}, \quad \mathbf{U} = \begin{bmatrix} 1 & 0 & -2 & 3 \\ & 1 & 3 & 1 \\ & & 1 & -2 \\ & & & 1 \end{bmatrix}.$$

Note: To illustrate the procedure, an example has been taken in which the elements of \mathbf{L} and \mathbf{U} are integral. This simplifies the numerical work, but is not important in principle.

Other methods of carrying out the factorization $\mathbf{A} = \mathbf{LU}$ and for solving the resulting equations have been proposed.† Some of these methods are designed to reduce the number of intermediate quantities which have to be written down in the course of the calculation; this is a matter of some importance in the treatment of large matrices. On the other hand, in methods designed with this object, too little is written down to indicate how the numbers which are written down were obtained, and this makes diagnosis of mistakes difficult; the use of row and column sums gives a good indication of freedom from mistakes when none have been made, but the checks are too few to be of much help in identifying or locating a mistake if one is indicated.

The absence of intermediate results to make it clear how each number in the course of the calculation is obtained means that the computer must depend on his memory for the required sequence of operations. This sequence is slightly different for every one of the elements of \mathbf{L} and of \mathbf{U} , and though it might become familiar enough to anyone who had much work of this kind to do, the writer's experience is that such a method is too complicated to be satisfactory for occasional use; it seems more suitable for the professional expert and the specialist rather than for the occasional user. In Wilkes's process given above, more may be written down, but the sequence of numerical operations is more apparent, and so easier to follow and to remember.

† See, for example, L. Fox, *Journ. Roy. Stat. Soc., Ser. B*, **12** (1952), 120.

8.41. Inverse of a matrix by Choleski's method

The inverse of a lower triangular matrix \mathbf{L} can be found by taking the identity $\mathbf{L}\mathbf{L}^{-1} = \mathbf{I}$, or in expanded form, with \mathbf{C} for \mathbf{L}^{-1}

$$\begin{bmatrix} l_{11} & 0 & 0 & 0 & . & . & . \\ l_{21} & l_{22} & 0 & 0 & . & . & . \\ l_{31} & l_{32} & l_{33} & 0 & . & . & . \\ . & . & . & . & . & . & . \end{bmatrix} \begin{bmatrix} c_{11} & c_{12} & c_{13} & . & . & . \\ c_{21} & c_{22} & c_{23} & . & . & . \\ c_{31} & c_{32} & c_{33} & . & . & . \\ . & . & . & . & . & . \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & . & . & . \\ 0 & 1 & 0 & 0 & . & . & . \\ 0 & 0 & 1 & 0 & . & . & . \\ . & . & . & . & . & . & . \end{bmatrix},$$

and working through the unit matrix column by column.

From the j th column of the unit matrix we have

$$\begin{aligned} l_{11}c_{1j} &= \delta_{1j} \\ l_{21}c_{1j} + l_{22}c_{2j} &= \delta_{2j} \\ l_{31}c_{1j} + l_{32}c_{2j} + l_{33}c_{3j} &= \delta_{3j} \\ . & \end{aligned}$$

(where $\delta_{ij} = 1$ if $j = i$ and $= 0$ if $j \neq i$). This is a set of equations of the form (8.20); the solution can be obtained by the 'division' procedure already explained for such a set of equations; this has to be done for each value of j . Since $\delta_{ij} = 0$ for $i \neq j$, it follows that $c_{ij} = 0$ for $i < j$, that is, that \mathbf{L}^{-1} is also lower triangular. This can also be seen from formulae (8.22), from which it follows that if $\mathbf{Ly} = \mathbf{b}$, then y_j is a linear combination of the b_i 's for $i \leq j$ only. The inverse of an upper triangular matrix can be found in a similar way, starting from the *last* column of the unit matrix.

If $\mathbf{A} = \mathbf{LU}$,

then $\mathbf{A}^{-1} = \mathbf{U}^{-1}\mathbf{L}^{-1}$,

so that a matrix can be inverted by factorizing it as a product \mathbf{LU} , inverting \mathbf{U} and \mathbf{L} separately, and finally multiplying them to form \mathbf{A}^{-1} .

8.5. Relaxation method

An indirect method which is very powerful in some cases is one called the 'relaxation' method. It was originally developed by Southwell† for application in problems of structural engineering, and some of the terminology of the method is derived from this particular application. But its range of application is much wider.

† R. V. Southwell, *Proc. Roy. Soc. A*, **151** (1935), 56; see also *ibid.* **184** (1945), 253; *Relaxation Methods in Engineering Science* (Oxford, 1940).

For any approximation $x_j = \xi_j$ to the solution of equations (8.1), the 'residuals' of the equations are defined by (8.2). For the solution of the equations, all the residuals are 0. In the relaxation method attention is concentrated on the residuals, and the method consists in making changes in the x 's in a systematic manner so as to reduce the magnitudes of the residuals to negligible amounts.

The first process is to draw up an 'operations table' giving the change of each residual for a unit change of each single x_j . Then a set of initial values of x_j is taken and the residuals calculated, and changes of the x_j 's then made in such a way as to decrease the residuals; the steps of this part of the work are recorded in a 'relaxation table' in which the *changes* of the x 's and the resulting *total* residuals are recorded.

Example: To find, correct to two decimal places, the solution of the equations

$$\left. \begin{aligned} 9x_1 - 2x_2 + x_3 &= 50 \\ x_1 + 5x_2 - 3x_3 &= 18 \\ -2x_1 + 2x_2 + 7x_3 &= 19 \end{aligned} \right\}. \quad (8.26)$$

For the first equation the residual R_1 , for any trial values of x_1, x_2, x_3 is

$$R_1 = 9x_1 - 2x_2 + x_3 - 50.$$

Hence for a change $\Delta x_1 = 1$ of x_1 alone, the change of R_1 is $\Delta R_1 = 9$; similarly for a change $\Delta x_2 = 1$ of x_2 alone, the change of R_1 is $\Delta R_1 = -2$, and for a change $\Delta x_3 = 1$ of x_3 alone, the change of R_1 is $\Delta R_1 = +1$. These values of ΔR_1 are entered in the *first column* of the operations table.

Similarly, the residual for the second equation is

$$R_2 = x_1 + 5x_2 - 3x_3 - 18,$$

and for the same changes $\Delta x_1, \Delta x_2, \Delta x_3$, the changes in R_2 are $\Delta R_2 = +1, +5$, and -3 respectively, and so on. Thus in this arrangement of the work, the matrix of these entries in the operations table is the transpose of the matrix of the coefficients in the equation. The working is shown on the opposite page.

Notes: (i) In this example, a change Δx_1 of x_1 affects R_1 mainly, and R_2, R_3 to a smaller extent, and similarly for changes Δx_2 and Δx_3 . A change Δx_1 , made in such a way as to reduce $|R_1|$ considerably, is called a 'relaxation' of x_1 , and the relaxation process consists of making a sequence of such relaxations. There is clearly no point in choosing Δx_1 so as to make R_1 exactly zero, since R_1 will be affected by subsequent relaxations Δx_2 and Δx_3 . It will usually be adequate to take single-digit numbers for the relaxations; this greatly lightens the numerical work, and makes it possible to carry out the greater part of it mentally and speedily.

(ii) The relaxation table begins with any trial set of values of x_1, x_2, x_3 and the corresponding residuals. The simplest first trial set is $x_1 = x_2 = x_3 = 0$, and the residuals are then just the negatives of the constant terms in the equations.

In this case the residual R_1 is the greatest, and a large part of this can be removed, with smaller changes in the other residuals, by a relaxation $\Delta x_1 = 5$. The resulting changes in all the residuals are given by multiplying line (1) of the operations table by 5. In the example on p. 187, these changes are shown in brackets; they would not be written down in actual working; each would be evaluated mentally and

| | Δx_1 | Δx_2 | Δx_3 | ΔR_1 | ΔR_2 | ΔR_3 | Notes |
|------------|----------------|----------------|----------------|--------------|--------------|--------------|---------------------|
| Operations | 1 | 0 | 0 | 9 | 1 | -2 | Line (i) |
| table | 0 | 1 | 0 | -2 | 5 | 2 | Line (ii) |
| | 0 | 0 | 1 | 1 | -3 | 7 | Line (iii) |
| <hr/> | | | | | | | |
| | $x_1 = 0$ | $x_2 = 0$ | $x_3 = 0$ | R_1 -50 | R_2 -18 | R_3 -19 | |
| Relaxation | 5 | (0) | (0) | (45) | (5) | (-10) | Line (i) $\times 5$ |
| table | | | | -5 | -13 | -29 | |
| | (0) | (0) | 4 | -1 | -25 | -1 | |
| | (0) | 5 | (0) | -11 | 0 | 9 | |
| | 1 | (0) | (0) | -2 | 1 | 7 | |
| | (0) | (0) | -1 | -3 | 4 | 0 | |
| | (0) | -1 | (0) | -1 | -1 | -2 | |
| <hr/> | | | | | | | |
| | $x_1 = 6$ | $x_2 = 4$ | $x_3 = 3$ | -1 | -1 | -2 | Check |
| <hr/> | | | | | | | |
| | $\times 10$ | | | | | | |
| | $10x_1 = 60$ | $10x_2 = 40$ | $10x_3 = 30$ | -10 | -10 | -20 | |
| | | | 3 | -7 | -19 | 1 | |
| | | 4 | | -15 | 1 | 9 | |
| | 2 | | | 3 | 3 | 5 | |
| | | | -1 | 2 | 6 | -2 | |
| | | -1 | | 4 | 1 | -4 | |
| | -1 | | | -5 | 0 | -2 | |
| <hr/> | | | | | | | |
| | $10x_1 = 61$ | $10x_2 = 43$ | $10x_3 = 32$ | -5 | 0 | -2 | Check |
| <hr/> | | | | | | | |
| | $\times 10$ | | | | | | |
| | 610 | 430 | 320 | -50 | 0 | -20 | |
| | 5 | | | -5 | 5 | -30 | |
| | | | 4 | -1 | -7 | -2 | |
| | | 1 | | -3 | -2 | 0 | |
| <hr/> | | | | | | | |
| | $100x_1 = 615$ | $100x_2 = 431$ | $100x_3 = 324$ | -3 | -2 | 0 | Check |
| <hr/> | | | | | | | |
| | $\times 10$ | | | | | | |
| | 6150 | 4310 | 3240 | -30 | -20 | 0 | |
| | | 4 | | -38 | 0 | 8 | |
| | 4 | | | -2 | 4 | 0 | |
| | | -1 | | 0 | -1 | -2 | |
| <hr/> | | | | | | | |
| | 6154 | 4313 | 3240 | 0 | -1 | -2 | Check |

added to the previous value of the corresponding residual, and only the new residual would be written down. For example, the value $\Delta R_1 = 9$ in the operations table, multiplied by $\Delta x_1 = 5$, gives 45, which added to the old $R_1 (= -50)$ gives the new $R_1 (= -5)$ and only this is written down. In subsequent lines in the relaxation table the only entries in the R_j columns are those which would be written down in actual working. Similarly, the bracketed zeros in the Δx_j columns might be omitted, as they have been later in the relaxation table.

After the first relaxation $\Delta x_1 = 5$ the largest residual is $R_3 = -29$ and most of this is removed by a further relaxation $\Delta x_3 = 4$; then R_2 is greatest and can be

reduced to zero by $\Delta x_2 = 5$. The contributions to R_1 from these relaxations Δx_2 and Δx_3 have been such that $|R_1|$ can be reduced substantially by a further relaxation $\Delta x_1 = 1$, and the procedure continues until the residuals are so small that they cannot be improved without making relaxations of magnitude smaller than unity.

At this stage it is convenient to avoid decimal points by multiplying the entries in the relaxation table by 10. But before doing this it is advisable to check that the calculation has been carried out correctly so far. This is done by adding up the changes Δx_1 and adding the result to the initial value taken (here zero), and similarly for x_2, x_3 . The resulting values of x_1, x_2, x_3 are then substituted in the equations and the residuals calculated; they should agree with the last line so far obtained in the operations table. If they do not, there is no need to go back and look for the mistake; these values of x_1, x_2, x_3 form as good a set of trial values as the set $(0, 0, 0)$ actually used—indeed they are probably much better—and can be used as the starting values of a further calculation.

(iii) When a set of small residuals has been obtained *and checked*, the values of x and the residuals can be multiplied by 10 and the process continued, and this operation can be repeated as often as required. In this example it is repeated until it gives the result

$$1000(x_1, x_2, x_3) = (6154, 4313, 3240)$$

or, to two decimals, $(x_1, x_2, x_3) = (6.15, 4.31, 3.24)$.

It will be seen that although the number of operations may be large, each is very simple and, except for the checking operations, each involves only numbers of two digits, and often only of one digit.

8.51. Group relaxations

There is no need to restrict oneself to relaxations of the variables singly. In this example a relaxation Δx_2 makes appreciable contributions to R_1 and R_3 , and the process of reducing the residuals would clearly be quicker and easier if we could make correlated changes in the variables in such a way as to affect only one residual considerably. It is often possible by trial to find linear combinations of the variables with this property. In this example we have:

| | Δx_1 | Δx_2 | Δx_3 | ΔR_1 | ΔR_2 | ΔR_3 |
|-------|--------------|--------------|--------------|--------------|--------------|--------------|
| X_3 | 0 | 1 | 2 | 0 | -1 | 16 |
| | 0 | 4 | -1 | -9 | 23 | 1 |
| X_2 | 1 | 4 | -1 | 0 | 24 | -1 |

A multiple of X_3 can clearly be used to reduce R_3 without affecting R_1 and with only quite a small effect on R_2 , and similarly a multiple of X_2 can be used to reduce R_2 . Relaxations which are multiples of X_2 or X_3 are called 'group relaxations' by Southwell. Use of them corresponds to making a linear transformation of the variables so that in the transformed variables the non-diagonal coefficients in the equations are small compared with the diagonal coefficients. But we do not have to carry out the transformation formally by introducing the new variables and

expressing the equations in terms of them. The process of carrying out the numerical work makes this transformation for us.

Use of these group relaxations does not, of course, preclude us from using relaxations of a single variable if the values of the residuals indicate that such a procedure would be appropriate.

Example

| | Δx_1 | Δx_2 | Δx_3 | ΔR_1 | ΔR_2 | ΔR_3 |
|------------|--------------|--------------|--------------|--------------------------|--------------|--------------|
| Operations | 1 | 0 | 0 | 9 | 1 | -2 |
| table | 0 | 1 | 0 | -2 | 5 | 2 |
| | 0 | 0 | 1 | 1 | -3 | 7 |
| <hr/> | | | | | | |
| X_2 | 1 | 4 | -1 | 0 | 24 | -1 |
| X_3 | 0 | 1 | 2 | 0 | -1 | 16 |
| <hr/> | | | | | | |
| Relaxation | x_1 | x_2 | x_3 | R_1 | R_2 | R_3 |
| table | 0 | 0 | 0 | -50 | -18 | -19 |
| | 5 | | | -5 | -13 | -29 |
| $2X_3$ | | 2 | 4 | -5 | -15 | 3 |
| X_2 | 1 | 4 | -1 | -5 | 9 | 2 |
| <hr/> | | | | | | |
| | 6 | 6 | 3 | -5 | 9 | 2 |
| <hr/> | | | | | | |
| | | | $\times 10$ | | | |
| | 60 | 60 | 30 | -50 | 90 | 20 |
| $-4X_2$ | -4 | -16 | 4 | -50 | -6 | 24 |
| | 5 | | | -5 | -1 | 14 |
| $-X_3$ | | -1 | -2 | -5 | 0 | -2 |
| <hr/> | | | | | | |
| | 61 | 43 | 32 | -5 | 0 | -2 |
| <hr/> | | | | | | |
| | | | $\times 10$ | | | |
| | 610 | 430 | 320 | -50 | 0 | -20 |
| | 6 | | | 4 | 6 | -32 |
| $2X_3$ | | 2 | 4 | 4 | 4 | 0 |
| <hr/> | | | | | | |
| | 616 | 432 | 324 | 4 | 4 | 0 |
| <hr/> | | | | | | |
| | | | $\times 10$ | | | |
| | 6160 | 4320 | 3240 | 40 | 40 | 0 |
| $-2X_2$ | -2 | -8 | 2 | 40 | -8 | 2 |
| | -4 | | | 4 | -12 | 10 |
| | | | -2 | 2 | -6 | -4 |
| | | 1 | | 0 | -1 | -2 |
| <hr/> | | | | | | |
| | 6154 | 4313 | 3240 | 0 | -1 | -2 |
| <hr/> | | | | | | |
| | 6.154 | 4.313 | 3.240 | Solution (to 3 decimals) | | |

8.52. Use and limitations of the relaxation method

Temple† has shown that if the simultaneous equations are derived from a positive definite quadratic form, then the relaxation process

† G. Temple, *Proc. Roy. Soc. A*, **169** (1938), 476.

formally converges. However, this is neither a necessary nor a sufficient condition for it to be a practicable process in actual numerical work; it is quite impracticable for the ill-conditioned equations (8.6) which can be derived from a positive definite quadratic form, whereas it is quite practicable for equations (8.26) which are not so derivable. What matters much more is that it should be easy to find a set of relaxations, either of individual variables or group relaxations, each of which has a considerably greater effect on one of the residuals than on any other.

Although the relaxation method is one of successive approximation, there is no limit in principle to the accuracy to which the solution can be taken. In the structural engineering context for which the method was originally devised by Southwell† there is no significance in carrying the solution beyond a certain limited degree of accuracy, and its first presentation in this context appears to have given the impression that it is fundamentally approximate in character; but it is no more so than any other numerical process which is subject to rounding errors. When it is used in contexts in which no approximation is involved in the equations to which it is applied, it may be possible, and significant, to carry the calculation through to a relatively high accuracy; and because of the simplicity of the process this may be the easiest way of deriving results to such accuracy.

For example, the recurrence relation for the Bessel functions

$$J_{n+1}(x) - (2n/x)J_n(x) + J_{n-1}(x) = 0$$

provides a set of simultaneous linear equations for $J_n(x)$ as a function of n for given x ; and given, say, $J_n(x)$ for $n = N \geq x$ and the condition $J_n(x) \rightarrow 0$ as $n \rightarrow \infty$, it is possible to solve these equations to any accuracy required by an application of the relaxation process. The process is a simple and quick one, and Fox‡ has recorded that starting only from $J_{10}(10)$, he has obtained eighteen-decimal values of $J_n(10)$, for values of n from 11 to the value ($n = 37$) at which $J_n(10) < 10^{-18}$, without difficulty. This process is most effective for calculating $J_n(x)$ for $n > x$, which is just the range over which the use of the recurrence relation to evaluate $J_{n+1}(x)$ from $J_n(x)$ and $J_{n-1}(x)$ becomes unsatisfactory (see § 11.3).

In one particular application, namely to the solution of ordinary differential equations with two-point boundary conditions‡ or of partial differential equations of elliptic type (§ 10.6), the relaxation process is

† R. V. Southwell, *Proc. Roy. Soc. A*, **151** (1935), 56.

‡ L. Fox, *Proc. Camb. Phil. Soc.* **45** (1949), 50.

combined with the use of finite-difference approximations to derivatives, and this is perhaps part of the reason why the relaxation process has come to be regarded as essentially approximate. But this is a mistaken idea. The approximation here is in the reduction of the differential equation to a set of simultaneous algebraic equations by use of finite differences. This approximation is involved whatever numerical process is used for the evaluation of a solution of these simultaneous equations, and the relaxation process, if used, is quite distinct from this approximation.

If a number of sets of equations with the same set of coefficients but with different values of (b_1, \dots, b_n) have to be solved, no advantage can be taken of this in the relaxation process, and for this reason this process is not well adapted to the inversion of a matrix.

The main difficulty likely to arise in using the process is slow convergence, and this will be most likely to occur when many of the coefficients in the equations are of the same order of magnitude, so that a relaxation of one of the unknowns makes changes of similar magnitude in a number of residuals. This situation is likely to occur when the equations are ill-conditioned, but is not necessarily a symptom of this condition.

8.6. Linear differential equations and linear simultaneous equations

Consider the linear differential equation

$$y'' = f(x)y + g(x) \quad (8.27)$$

with two-point boundary conditions, $y = y_0$ at $x = x_0$, $y = y_n$ at $x = x_n = x_0 + n(\delta x)$. To the approximation in which $(\delta x)^2 y''$ can be replaced by $\delta^2 y$, this differential equation is equivalent to the set of algebraic equations

$$\delta^2 y_j = (\delta x)^2 [f_j y_j + g_j],$$

$$\text{or} \quad y_{j+1} - [2 + (\delta x)^2 f_j] y_j + y_{j-1} = (\delta x)^2 g_j, \quad (8.28)$$

for $1 \leq j \leq n-1$, with y_0 and y_n given. This is a set of linear algebraic equations for $(n-1)$ unknowns, and could be solved numerically by any of the methods for the solution of such equations.

A convenient process in many cases is one that has been suggested and developed by Thomas and by Fox.† It is a version of the Choleski

† See L. Fox and H. H. Robertson, *Proceedings of a Symposium on Automatic Digital Computation*, N.P.L. 1953 (H.M.S.O., 1954), ch. 19, section on 'Boundary Value Problems'; also L. Fox, *The Numerical Solution of Two-point Boundary Problems in Ordinary Differential Equations* (Clarendon Press, 1957), ch. 3, § 29. As far as I am aware, Thomas's work on this process has not been published; see the Preface to Fox's book.

method (§ 8.4), which in this case is very simple because of the specially simple forms of the matrices concerned.

Equations (8.28), arranged with the known quantities on the right-hand side, are

$$\left. \begin{aligned} -[2 + (\delta x)^2 f_1]y_1 + y_2 &= -y_0 + (\delta x)^2 g_1 \\ y_{j-1} - [2 + (\delta x)^2 f_j]y_j + y_{j+1} &= (\delta x)^2 g_j \quad (2 \leq j \leq n-2) \\ y_{n-2} - [2 + (\delta x)^2 f_{n-1}]y_{n-1} &= -y_n + (\delta x)^2 g_{n-1} \end{aligned} \right\} \quad (8.29)$$

For short, let $\phi_j = (\delta x)^2 f_j$, and

$c_1 = -y_0 + (\delta x)^2 g_1$, $c_j = (\delta x)^2 g_j$ ($2 \leq j \leq n-2$), $c_{n-1} = -y_n + (\delta x)^2 g_{n-1}$;

also let $\mathbf{y} = (y_1, y_2, \dots, y_n)$, $\mathbf{c} = (c_1, c_2, \dots, c_n)$. Then equations (8.29), in matrix form, are

$$\mathbf{A}\mathbf{y} = \mathbf{c}, \quad (8.30)$$

where \mathbf{A} is the matrix

$$\mathbf{A} = \begin{bmatrix} -(2+\phi_1) & 1 & 0 & 0 & \cdot & \cdot & 0 & 0 & 0 \\ 1 & -(2+\phi_2) & 1 & 0 & \cdot & \cdot & 0 & 0 & 0 \\ 0 & 1 & -(2+\phi_3) & 1 & \cdot & \cdot & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & 0 & \cdot & \cdot & 1 & -(2+\phi_{n-2}) & 1 \\ 0 & 0 & 0 & 0 & \cdot & \cdot & 0 & 1 & -(2+\phi_{n-1}) \end{bmatrix}.$$

It can easily be verified that for this matrix \mathbf{A} the lower and upper triangular matrices of § 8.4 are

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & 0 & \cdot & \cdot & 0 & 0 & 0 \\ -l_1 & 1 & 0 & \cdot & \cdot & 0 & 0 & 0 \\ 0 & -l_2 & 1 & \cdot & \cdot & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \cdot & \cdot & -l_{n-3} & 1 & 0 \\ 0 & 0 & 0 & \cdot & \cdot & 0 & -l_{n-2} & 1 \end{bmatrix}, \quad (8.31)$$

$$\mathbf{U} = \begin{bmatrix} -1/l_1 & 1 & 0 & \cdot & \cdot & 0 & 0 & 0 \\ 0 & -1/l_2 & 1 & \cdot & \cdot & 0 & 0 & 0 \\ 0 & 0 & -1/l_3 & \cdot & \cdot & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \cdot & \cdot & -1/l_{n-2} & 1 & 0 \\ 0 & 0 & 0 & \cdot & \cdot & 0 & -1/l_{n-1} & 0 \end{bmatrix},$$

where

$$1/l_1 = (2+\phi_1), \quad l_{j-1} + (1/l_j) = (2+\phi_j) \quad (2 \leq j \leq n-1)$$

so that

$$l_1 = 1/(2+\phi_1), \quad l_j = 1/(2+\phi_j - l_{j-1}) \quad (2 \leq j \leq n-1). \quad (8.32)$$

From these relations the l_j 's can be evaluated in succession, in order of increasing j .

Also equation (8.30) becomes $\mathbf{LUy} = \mathbf{c}$, so that if $\mathbf{z} = -\mathbf{Uy}$, then $\mathbf{Lz} = -\mathbf{c}$, and for the matrix \mathbf{L} given by (8.31) it follows that

$$z_1 = -c_1, \quad z_j = -c_j + l_{j-1}z_{j-1} \quad (2 \leq j \leq n-1), \quad (8.33)$$

from which the z_j 's can be evaluated in succession, in the order of j increasing; then from the equation $\mathbf{z} = -\mathbf{Uy}$ it follows that

$$y_{n-1} = l_{n-1}z_{n-1}, \quad y_j = l_j(y_{j+1} + z_j) \quad (1 \leq j \leq n-2) \quad (8.34)$$

from which the y_j 's can be evaluated in succession, in the order of j decreasing.

The process is a direct one, and, if solutions of (8.30) are required for the same function $f(x)$ and several different vectors \mathbf{c} , the matrix \mathbf{A} , and so the values of the l_j 's, are the same for all vectors \mathbf{c} , so the evaluation of the l_j 's only has to be done once. This may be necessary if it is required to improve the approximation to the solution of the differential equation (8.27) by including higher orders of differences in the replacement of derivatives by differences. The process is easily adaptable to second-order linear differential equations containing a first-derivative term, and was indeed formulated by Thomas and by Fox for this more general case, but it is particularly simple for equations like (8.27) for which the first derivative is absent. Also higher-order differences can be included in the replacement of y'' by finite differences, as follows.

The set of equations (8.28) has been obtained by replacing the derivative in (8.27) by the simplest finite-difference approximation to it. In the solution of the equations in the approximate form (8.28), y'' is never evaluated; so in taking higher orders of differences into account, it is more convenient to use those of y than those of y'' . The next approximation to y_j'' is given by

$$(\delta x)^2 y_j'' = \delta^2 y_j - \frac{1}{12} \delta^4 y_j;$$

use of this approximation in (8.27) gives

$$y_{j+1} - [2 + (\delta x)^2 f_j] y_j + y_{j-1} = (\delta x)^2 g_j + \frac{1}{12} \delta^4 y_j. \quad (8.35)$$

This set of equations should be solved by an iterative process, the values of $\delta^4 y$ on the right-hand side in one iteration being obtained from the results of the previous iteration; higher orders of differences of y can be included on the right-hand side of (8.35) if appreciable. One should *not* try to solve it by expressing $\delta^4 y_j$ in terms of function values and treating (8.35) as a recurrence relation between five successive values of y ; this would be equivalent to replacing equation (8.27) by a fourth-order equation and would probably introduce spurious 'solutions'.

The terms $\frac{1}{12}\delta^4 y_j$ (and higher difference terms if appreciable) can be incorporated in the c_j 's.

Example: To find an approximation to the solution of the equation

$$y'' = x^2 y - 1 \quad (8.36)$$

with $y = 0$ at $x = \pm 4$.

From the symmetry of the equation and the boundary conditions, it follows that y is an even function of x . Hence we need only consider the range $x = 0$ to 4, and impose a condition of symmetry about $x = 0$.

Let us take $\delta x = \frac{1}{2}$ and reckon j from $j = 0$ at $x = 0$. Then equation (8.28) becomes

$$y_{j+1} - (2 + \frac{1}{16}j^2)y_j + y_{j-1} = -\frac{1}{4} + \frac{1}{12}\delta^4 y_j \quad \text{for } 1 \leq j \leq 7, \quad (8.37)$$

and the condition of symmetry, $y_{-1} = y_1$, gives for $j = 0$ the equation

$$-y_0 + y_1 = -\frac{1}{8} + \frac{1}{24}\delta^4 y_0; \quad (8.38)$$

so that

$$\phi_j = \frac{1}{16}j^2, \quad l_0 = 1; \quad c_0 = -\frac{1}{8} + \frac{1}{24}\delta^4 y_0, \quad c_j = -\frac{1}{4} + \frac{1}{12}\delta^4 y_j \quad (1 \leq j \leq 7); \quad (8.39)$$

and equations (8.32), (8.33) apply for $1 \leq j \leq 7$.

The working can be arranged as shown on p. 195; some intermediate results, which it would not be necessary to write down if working with a desk machine, are included to show the sequence of operations.

In this working, the calculation is split into three sections by heavy lines. The first section, to the left of the first heavy line, is concerned with the calculation of the l_j 's. The values of $(2 + \phi_j)$ are filled in first from the formula for ϕ_j , then the l_j 's calculated in succession from formulae (8.32). The second section is concerned with a first approximation to y_j , neglecting the fourth-difference terms in (8.37), (8.38). The z_j 's are calculated first, by working downwards through the columns to the left of the thin line; each value of z_j is multiplied by the value of l_j in the *same* line, and the product is written in the *next lower* line in the column headed $l_{j-1}z_{j-1}$, and added to the $(-c_j)$ in that line to give the next z_j .

The next three columns are concerned with the evaluation of y_j starting from $y_8 = 0$ and working *upwards*; as each y_j is calculated it is entered in the *next higher* line in the column headed y_{j+1} and added to the z_j in that line; the sum is then multiplied by l_j in that line.

The next three columns are concerned with the evaluation of the $\delta^4 y$ terms in formula (8.39). The third section of the calculation is a repetition of the procedure of the second section with the $\delta^4 y_j$ contributions to the c_j 's included.

Notes: (i) The symmetry of the solution about $y = 0$ has been used in evaluation $\delta^2 y_0$ and $\delta^4 y_0$.

(ii) Since y is required to be zero at $x = 4$ ($j = 8$), $y''(4) = -1$, so

$$\delta^2 y_8 = (\delta x)^2 y''(4) + O(\delta x)^4 = -0.250 \text{ approximately.}$$

This value (enclosed in brackets) has been used to give a value of $\delta^4 y_7$.

(iii) The calculation could be repeated with values of $\delta^4 y$ derived from the second iteration.

(iv) Smaller intervals (δx) should be used if greater accuracy in the results were required.

(v) This process may not be satisfactory when the function f in equation (8.27) is negative over a considerable range of x .

| j | l_j | $-c_0$ | z_0 | y_{j+1} | $z_j + y_{j+1}$ | y_j | $\delta^2 y_j$ | $\delta^4 y_j$ | $\frac{1}{2} \delta^4 y_0$ |
|-----|--------------|----------|-------------------|-----------|-----------------|-------|----------------|----------------|----------------------------|
| 0 | $2 + \phi_j$ | $+0.125$ | $= 0.125$ | 1.209 | 1.334 | 1.334 | -250 | 152 | +6 |
| 1 | $-l_{j+1}$ | $-c_j$ | $+l_{j-1}z_{j-1}$ | $= z_j$ | | | | | $\frac{1}{2} \delta^4 y_j$ |
| 2 | 2.0625 | 0.25 | $+0.125$ | $= 0.375$ | 1.285 | 1.209 | -174 | 75 | +6 |
| 3 | 2.25 | 0.25 | $+0.353$ | $= 0.603$ | 1.191 | 0.910 | -23 | -47 | -4 |
| 4 | 2.5625 | 0.25 | $+0.461$ | $= 0.711$ | 1.058 | 0.588 | +81 | -89 | -7 |
| 5 | 3.00 | 0.25 | $+0.395$ | $= 0.645$ | 0.847 | 0.347 | 96 | -44 | -4 |
| 6 | 3.5625 | 0.25 | $+0.264$ | $= 0.514$ | 0.638 | 0.202 | 67 | -10 | -1 |
| 7 | 4.25 | 0.25 | $+0.163$ | $= 0.413$ | 0.487 | 0.124 | 28 | -13 | -1 |
| 8 | 5.0625 | 0.25 | $+0.105$ | $= 0.355$ | 0.353 | 0.074 | -24 | -174 | -15 |
| | | | | | 0 | 0 | (-250) | | |

| j | l_j | $-c_0$ | z_0 | y_{j+1} | $z_j + y_{j+1}$ | y_j | $\delta^2 y_j$ | $\delta^4 y_j$ |
|-----|--------|--------|-------------------|-----------|-----------------|-------|----------------|----------------|
| 0 | 1 | 0.119 | 0.119 | 1.194 | 1.313 | 1.313 | -238 | 158 |
| 1 | 0.9412 | $-c_j$ | $+l_{j-1}z_{j-1}$ | $= z_j$ | | | | |
| 2 | 0.7641 | 0.244 | $+0.119$ | $= 0.363$ | 0.906 | 1.194 | -169 | 62 |
| 3 | 0.5560 | 0.254 | $+0.342$ | $= 0.586$ | 0.590 | 0.906 | -28 | -28 |
| 4 | 0.4092 | 0.257 | $+0.455$ | $= 0.712$ | 0.349 | 0.590 | +75 | -82 |
| 5 | 0.3171 | 0.254 | $+0.396$ | $= 0.650$ | 0.204 | 0.349 | 96 | -51 |
| 6 | 0.2543 | 0.251 | $+0.266$ | $= 0.517$ | 0.125 | 0.204 | 66 | -5 |
| 7 | 0.2080 | 0.251 | $+0.164$ | $= 0.415$ | 0.077 | 0.125 | 31 | -15 |
| 8 | | 0.265 | $+0.105$ | $= 0.370$ | 0 | 0.077 | -29 | -171 |
| | | | | | 0 | 0 | (-250) | |

Entries in columns headed y_j calculated from formulae (8.34).

If a boundary condition involves the derivative y' , it can best be handled by imagining the range of x extended one interval beyond the actual boundary, and eliminating the value of y at the virtual external point. Suppose, for example, that the boundary condition at $x = 0$ is

$$\alpha y'_0 + \beta y_0 = \gamma,$$

α , β , and γ being given. y_0 is not now known, and an extra equation is required to determine it.

Let suffix -1 refer to the external point. Then for $x = x_0$ equation (8.28) becomes

$$y_1 - [2 + (\delta x)^2 f_0] y_0 + y_{-1} = (\delta x)^2 g_0,$$

and, to the same order of accuracy, we have

$$y'_0 = (y_1 - y_{-1}) / 2\delta x.$$

Elimination of y'_0 and y_{-1} between these three equations gives a relation between y_0 and y_1 which takes the place of the first of equations (8.29); and the range of j to which the second of equations (8.29) applies now extends to $j = 1$.

A similar treatment applies to a boundary condition at x_n involving a derivative, and to conditions at a point x_k , between x_0 and x_n , at which $f(x)$ is discontinuous but y and y' continuous, or at which δx changes.

An alternative procedure for the solution of equations (8.29) is a relaxation process.† It is best to start such a process with quite a coarse subdivision of the range over which the solution is wanted (such as $\delta x = 1$ in the above example) and to divide it further as the approximation improves. Beyond a certain stage of the subdivision, depending on the behaviour of y as a function of x and the accuracy required in the solution, it is often better, as suggested by Fox,† to keep some higher differences in the replacement of y'' by finite differences.

8.7. Characteristic values and vectors of a matrix

The 'characteristic values' (also called 'latent roots') λ of a matrix are those numbers for which the system of equations

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x} \tag{8.40}$$

has a solution other than $\mathbf{x} \equiv 0$, and the solution \mathbf{x} in such a case is called a 'characteristic vector' (or 'latent vector') of the matrix or a 'characteristic solution' of the equations.

† L. Fox, *Proc. Camb. Phil. Soc.* **45** (1949), 50, and *The Numerical Solution of Two-point Boundary Problems in Ordinary Differential Equations* (Clarendon Press, 1957), ch. 3, §§ 20–25.

Written out, the system of equations (8.40) is

$$\left. \begin{aligned} (a_{11}-\lambda)x_1 + a_{12}x_2 + a_{13}x_3 + \dots &= 0 \\ a_{21}x_1 + (a_{22}-\lambda)x_2 + a_{23}x_3 + \dots &= 0 \\ a_{31}x_1 + a_{32}x_2 + (a_{33}-\lambda)x_3 + \dots &= 0 \\ \vdots &\vdots \end{aligned} \right\}, \quad (8.41)$$

and the condition that this set of equations should have a non-trivial solution is that the determinant of its coefficients should be zero. An algebraic equation of the n th degree in λ can be obtained by multiplying out the determinant, though for $n > 2$ it is seldom that this provides the easiest or quickest way of evaluating characteristic values numerically. But this formal equation gives three results which are useful in numerical work. The equation is

$$(-1)^n \left[\lambda^n - \left(\sum_j a_{jj} \right) \lambda^{n-1} + \dots \right] + D = 0, \quad (8.42)$$

where D is the determinant of the matrix. Hence

- (i) the equation for λ has n roots (multiple roots, if any, being counted according to their multiplicity);
- (ii) the sum of the roots is the sum of the diagonal elements of the matrix;
- (iii) the product of the roots is the determinant of the matrix.

The characteristic values λ , arranged in order of decreasing $|\lambda|$, will be written $\lambda^{(1)}, \lambda^{(2)}, \lambda^{(3)}, \dots$, and the solution x_j for $\lambda = \lambda^{(p)}$ will be written $x_j^{(p)}$.

Since equations (8.41) are homogeneous, each solution $\mathbf{x}^{(p)}$ is undetermined to the extent of an arbitrary multiplying factor. In the formal treatment of these equations it is convenient to regard this multiplying constant as chosen so that $\sum_j (x_j^{(p)})^2 = 1$; such a solution is called 'normalized'. But for numerical work it is usually more convenient to take it so that the greatest in magnitude of the x_j 's is unity.

A commonly occurring type of matrix is a symmetrical matrix, and symmetrical matrices have several properties which are useful in numerical calculations. The characteristic values λ are all real, and the characteristic vectors for different values of λ are orthogonal, that is to say, if $\lambda^{(p)} \neq \lambda^{(q)}$, then

$$\sum_j x_j^{(p)} x_j^{(q)} = 0; \quad (8.43)$$

and if any value of λ is an m -fold root of the equation (8.42), then m mutually orthogonal solutions x_j can be found for this value of λ . Hence it is no restriction to take all the characteristic vectors as mutually orthogonal, and it will be assumed that they are so taken.

Further, the behaviour of the quantity

$$\Lambda = \left(\sum_{jk} x_j a_{jk} x_k \right) / \left(\sum_j x_j^2 \right) \quad (8.44)$$

as a function of the vector \mathbf{x} has a property which makes it useful in the approximate calculation of characteristic values. In matrix notation this quantity can be written

$$\Lambda = (\mathbf{x}' \mathbf{A} \mathbf{x}) / (\mathbf{x}' \mathbf{I} \mathbf{x}), \quad (8.45)$$

where \mathbf{I} is the unit matrix, and \mathbf{x}' is the row vector corresponding to the column vector \mathbf{x} .

Consider first the result of evaluating Λ for one of the characteristic vectors of the matrix, say for $\mathbf{x} = \mathbf{x}^{(p)}$. From the definition of a characteristic vector, it follows that $\mathbf{A} \mathbf{x}^{(p)} = \lambda^{(p)} \mathbf{x}^{(p)}$, and hence that the value of Λ obtained is just $\Lambda = \lambda^{(p)}$, the corresponding characteristic value. That is

$$\lambda^{(p)} = (\mathbf{x}^{(p)'} \mathbf{A} \mathbf{x}^{(p)}) / (\mathbf{x}^{(p)'} \mathbf{I} \mathbf{x}^{(p)}).$$

Now consider the value of Λ evaluated for a vector \mathbf{x} differing slightly from $\mathbf{x}^{(p)}$, say for $\mathbf{x} = \mathbf{x}^{(p)} + \boldsymbol{\xi}$. From (8.45), its difference from the value $\lambda^{(p)}$ is given by

$$(\Lambda - \lambda^{(p)}) (\mathbf{x}' \mathbf{I} \mathbf{x}) = \mathbf{x}' (\mathbf{A} - \lambda^{(p)} \mathbf{I}) \mathbf{x}. \quad (8.46)$$

Since $\mathbf{x}^{(p)}$ is a characteristic vector with characteristic value $\lambda^{(p)}$ it follows that $(\mathbf{A} - \lambda^{(p)} \mathbf{I}) \mathbf{x}^{(p)} = 0$, and hence

$$(\mathbf{A} - \lambda^{(p)} \mathbf{I}) \mathbf{x} = (\mathbf{A} - \lambda^{(p)} \mathbf{I}) \boldsymbol{\xi}. \quad (8.47)$$

Since \mathbf{A} is symmetrical, so is $\mathbf{A} - \lambda^{(p)} \mathbf{I}$; hence from (8.46) and (8.47)

$$\begin{aligned} (\Lambda - \lambda^{(p)}) (\mathbf{x}' \mathbf{I} \mathbf{x}) &= \mathbf{x}' (\mathbf{A} - \lambda^{(p)} \mathbf{I}) \boldsymbol{\xi} \\ &= \boldsymbol{\xi}' (\mathbf{A} - \lambda^{(p)} \mathbf{I}) \mathbf{x} \quad (\text{since } \mathbf{A} - \lambda^{(p)} \mathbf{I} \text{ is symmetrical}) \\ &= \boldsymbol{\xi}' (\mathbf{A} - \lambda^{(p)} \mathbf{I}) \boldsymbol{\xi} \quad (\text{by a second use of (8.47)}). \end{aligned}$$

Hence Λ differs from $\lambda^{(p)}$ by a quantity which is second-order in $\boldsymbol{\xi}$; in other words, the quantity Λ defined by (8.44) is stationary for small variations of the vector \mathbf{x} from a characteristic vector.

Hence from a fair approximation to a characteristic vector a relatively good approximation to a characteristic value can be obtained by evaluating formula (8.44), and from a good approximation to a characteristic vector a much better approximation to a characteristic value can be obtained. In application to the equations of vibrating systems of several degrees of freedom, characteristic values represent squares of frequencies of normal modes of vibration; an account of the use of formula (8.44) to determine characteristic values in this context, and

developments of it, is given in *Rayleigh's Principle*, by Temple and Bickley.†

It is sometimes convenient to express a symmetrical matrix in terms of its characteristic vectors and characteristic values. Let $\mathbf{x}'\mathbf{x}$ stand for the matrix whose (j, k) th element is $x_j x_k$. Then if the characteristic vectors $\mathbf{x}^{(p)}$ are normalized, the required expression is

$$\mathbf{A} = \sum_p \lambda^{(p)} \mathbf{x}^{(p)'} \mathbf{x}^{(p)};$$

if, as is often more convenient for numerical work, they are not normalized, we have

$$\mathbf{A} = \sum_p \left[\lambda^{(p)} \mathbf{x}^{(p)'} \mathbf{x}^{(p)} / \sum_j (x_j^{(p)})^2 \right]. \quad (8.48)$$

The characteristic vectors of the inverse \mathbf{A}^{-1} of a symmetrical matrix \mathbf{A} are the same as those of \mathbf{A} itself, and the characteristic values of \mathbf{A}^{-1} are the reciprocals of the corresponding characteristic values of \mathbf{A} , so that

$$\mathbf{A}^{-1} = \sum_p \left[(1/\lambda^{(p)}) \mathbf{x}^{(p)'} \mathbf{x}^{(p)} / \sum_j (x_j^{(p)})^2 \right]. \quad (8.49)$$

Formally, this provides one method of inverting a matrix. But it may be more convenient in numerical work to invert the matrix by some other process, and then use (8.49) to determine the *small* characteristic values of \mathbf{A} and their characteristic vectors.

8.71. Iterative method for evaluation of characteristic values and characteristic vectors of a symmetrical matrix

The characteristic value $\lambda^{(1)}$ of greatest magnitude can be found as follows. Take an arbitrary vector $\mathbf{x}_{(0)}$, with components $x_{(0)j}$, of which the greatest in magnitude is unity. Form $\mathbf{A}\mathbf{x}_{(0)}$ and express it as a multiple $\lambda_{(1)}$ of a vector $\mathbf{x}_{(1)}$ whose component of greatest value is unity. Then repeat the process with $\mathbf{x}_{(1)}$ in place of $\mathbf{x}_{(0)}$ to give a vector $\mathbf{x}_{(2)}$ and so on. That is, form a sequence of numbers $\lambda_{(m)}$ and of vectors $\mathbf{x}_{(m)}$ so that

$$\mathbf{A}\mathbf{x}_{(m-1)} = \lambda_{(m)} \mathbf{x}_{(m)},$$

where each $\lambda_{(m)}$ is chosen so that the component of $\mathbf{x}_{(m)}$ of greatest magnitude is unity. Then, unless the vector $\mathbf{x}_{(0)}$ happens to have been taken orthogonal to the characteristic vector $\mathbf{x}^{(1)}$ of the matrix,

$$\lambda_{(m)} \rightarrow \lambda^{(1)} \quad \text{and} \quad \mathbf{x}_{(m)} \rightarrow \mathbf{x}^{(1)} \quad \text{as } m \rightarrow \infty.$$

The process thus far is also applicable to non-symmetrical matrices.

The rate at which successive values of $\lambda_{(m)}$ ultimately tend to their limit depends on $|\lambda^{(1)}/\lambda^{(2)}|$, and is greater the greater the value of this

† G. Temple and W. G. Bickley, *Rayleigh's Principle* (Oxford, 1933).

ratio. If after the first few repetitions of the process of calculating $\lambda_{(m)}\mathbf{x}_{(m)} = \mathbf{A}\mathbf{x}_{(m-1)}$, the successive values of $\lambda_{(m)}$ seem to be tending to a limit only slowly, the reason may be that $|\lambda^{(1)}/\lambda^{(2)}| - 1$ is small. But it may be that $\mathbf{x}_{(0)}$ happens to have been taken nearly orthogonal to $\mathbf{x}^{(1)}$, and in case it has been so taken, it is as well to start the calculation again with another vector $\mathbf{x}_{(0)}$ roughly orthogonal to the one previously used.

Example: To find $\mathbf{x}^{(1)}$, $\lambda^{(1)}$ for the symmetrical matrix

$$\mathbf{A} = \begin{bmatrix} 23 & 11 & 1 \\ 11 & -3 & -2 \\ 1 & -2 & 1 \end{bmatrix}.$$

(This is the matrix of the coefficients of the equations considered in the examples of §§ 8.22 and 8.3.)

Starting with $\mathbf{x}_{(0)} = (1, 0, 0)$ we have

$$\begin{aligned} \mathbf{x}_{(0)} &= (1, 0, 0), & \mathbf{A}\mathbf{x}_{(0)} &= (-23, 11, 1) \\ & & &= -23(1, -0.48, -0.04), \\ \mathbf{x}_{(1)} &= (1, -0.48, -0.04), & \mathbf{A}\mathbf{x}_{(1)} &= (-28.32, 12.52, 1.92) \\ & & &= -28.32(1, -0.442, -0.068), \\ \mathbf{x}_{(2)} &= (1, -0.442, -0.068), & \mathbf{A}\mathbf{x}_{(2)} &= (-27.93, 12.462, 1.814) \\ & & &= -27.93(1, -0.4461, -0.0649), \\ \mathbf{x}_{(3)} &= (1, -0.4461, -0.0649), & \mathbf{A}\mathbf{x}_{(3)} &= (-27.971, 12.468, 1.827) \\ & & &= -27.971(1, -0.4458, -0.0653), \\ \mathbf{x}_{(4)} &= (1, -0.4458, -0.0653), & \mathbf{A}\mathbf{x}_{(4)} &= (-27.9691, 12.4680, 1.8263) \\ & & &= -27.9691(1, -0.44579, -0.06530). \end{aligned}$$

Hence, to three decimals in $\lambda^{(1)}$ and five in $\mathbf{x}^{(1)}$

$$\lambda^{(1)} = -27.969, \quad \mathbf{x}^{(1)} = (1, -0.44579, -0.06530). \quad (8.50)$$

Note: The number of figures in $\lambda_{(m)}$ and $\mathbf{x}_{(m)}$ can be kept small in the early stages and increased as the approximation to $\lambda^{(1)}$, $\mathbf{x}^{(1)}$ improves with repetition of the iterative process.

After determining $\lambda^{(1)}$ and $\mathbf{x}^{(1)}$, the characteristic value $\lambda^{(2)}$ of next greatest modulus and the corresponding characteristic vector can be found by repeating the procedure used for determining $\lambda^{(1)}$ and $\mathbf{x}^{(1)}$, with the modification that $\mathbf{x}_{(0)}$ and each $\mathbf{x}_{(m)}$ is constrained to be orthogonal to $\mathbf{x}^{(1)}$ before being multiplied by \mathbf{A} . That is, we form a sequence of numbers $\lambda_{(m)}$ and vectors $\mathbf{x}_{(m)}$ by the relation

$$\lambda_{(m)}\mathbf{x}_{(m)} = \mathbf{A}\mathbf{x}_{(m-1)} - \mu_{(m)}\mathbf{x}^{(1)},$$

$\mu_{(m)}$ being a number determined by the condition that $\mathbf{x}_{(m)}$ should be orthogonal to $\mathbf{x}^{(1)}$; $\lambda_{(m)}$, as before, is determined so that the component of $\mathbf{x}_{(m)}$ of greatest magnitude is unity.

If the work could be done exactly without rounding errors, then there would be no need to introduce the multipliers $\mu_{(m)}$; if $\mathbf{x}_{(0)}$ were taken

exactly orthogonal to $\mathbf{x}^{(1)}$, then each vector $\mathbf{A}\mathbf{x}_{(m)}$ would be orthogonal to $\mathbf{x}^{(1)}$. The rounded values of $\mathbf{A}\mathbf{x}_{(m-1)}$ will, however, contain a small multiple of $\mathbf{x}^{(1)}$, and if this is not removed, it will give rise to an error which increases with further repetition of the iterative process, so that ultimately $\mathbf{x}_{(m)}$ would tend to $\mathbf{x}^{(1)}$ and not to $\mathbf{x}^{(2)}$.

When $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ have been determined, the successive approximations to $\mathbf{x}^{(3)}$ are similarly constrained to be orthogonal both to $\mathbf{x}^{(1)}$ and to $\mathbf{x}^{(2)}$, and so on.

This process has the advantage that there is no loss of significant figures as the work proceeds. On the other hand, since it makes use of the orthogonal property of the characteristic vectors, this property cannot be used as a check, whereas it forms a valuable check when the characteristic vectors are determined independently of one another. Also it is necessary to determine each characteristic vector accurately before starting to calculate the next.

8.72. Richardson's purification process for characteristic vectors

In determining a characteristic vector $\mathbf{x}^{(a)}$ and corresponding characteristic value $\lambda^{(a)}$ by a method of successive approximation, it is an advantage to start with a trial approximation which is nearly orthogonal to any characteristic vectors about which something may be known from earlier stages of the calculation. One method of doing this has been outlined in § 8.71. Another process has been suggested by L. F. Richardson,[†] and has two advantages; first, it does not require accurate determination of other characteristic vectors, even those for which $|\lambda^{(p)}|$ is greater than $|\lambda^{(a)}|$ for the characteristic vector sought, and, secondly it enables the characteristic vectors to be calculated independently of one another so that the orthogonal property is available as a check.

Any vector \mathbf{x} (of n components) can be expressed as a linear combination of the characteristic vectors $\mathbf{x}^{(p)}$ of \mathbf{A} :

$$\mathbf{x} = b_1 \mathbf{x}^{(1)} + b_2 \mathbf{x}^{(2)} + \dots + b_n \mathbf{x}^{(n)} = \sum_p b_p \mathbf{x}^{(p)}. \quad (8.51)$$

Multiplication by $(\mathbf{A} - l\mathbf{I})$ gives

$$(\mathbf{A} - l\mathbf{I})\mathbf{x} = \sum_p (\lambda^{(p)} - l)b_p \mathbf{x}^{(p)}. \quad (8.52)$$

If one of the characteristic values, say $\lambda^{(r)}$, were known exactly, and l were taken to be $\lambda^{(r)}$, the coefficient of the corresponding term in (8.52)

[†] *Phil. Trans. Roy. Soc.* **242** (1950), 439.

would be zero; that is to say, $(\mathbf{A} - \lambda^{(r)}\mathbf{I})\mathbf{x}$ is orthogonal to $\mathbf{x}^{(r)}$. Similarly $(\mathbf{A} - \lambda^{(r)}\mathbf{I})(\mathbf{A} - \lambda^{(s)}\mathbf{I})\mathbf{x}$ is orthogonal to $\mathbf{x}^{(r)}$ and to $\mathbf{x}^{(s)}$, and so on.

Even if $\lambda^{(r)}$ is not known exactly, the coefficient of $\mathbf{x}^{(r)}$ in (8.52) will be relatively small provided that the value of l taken, say $l^{(r)}$, is such that $|\lambda^{(r)} - l^{(r)}|$ is substantially smaller than most of the other quantities $|\lambda^{(p)} - l^{(r)}|$ for $p \neq r$. Further, by repeating the multiplication by $(\mathbf{A} - l\mathbf{I})$ we have

$$(\mathbf{A} - l\mathbf{I})^m \mathbf{x} = \sum_p (\lambda^{(p)} - l)^m b_p \mathbf{x}^{(p)},$$

so that starting from any arbitrary vector \mathbf{x} , the vectors \mathbf{x} , $(\mathbf{A} - l^{(r)}\mathbf{I})\mathbf{x}$, $(\mathbf{A} - l^{(r)}\mathbf{I})^2\mathbf{x}$, ... are more and more nearly orthogonal to $\mathbf{x}^{(r)}$. Similarly if $l^{(r)}$, $l^{(s)}$ are approximations to $\lambda^{(r)}$, $\lambda^{(s)}$, and any vector \mathbf{x} is multiplied by $(\mathbf{A} - l^{(r)}\mathbf{I})(\mathbf{A} - l^{(s)}\mathbf{I})$, the result will be a vector nearly orthogonal to $\mathbf{x}^{(r)}$ and to $\mathbf{x}^{(s)}$, and so on.

If $\mathbf{x}^{(q)}$ is a characteristic vector to be determined, an arbitrary vector \mathbf{x} can be described as a mixture of the required vector $\mathbf{x}^{(q)}$ with 'impurities' in the form of multiples of the $\mathbf{x}^{(p)}$'s ($p \neq q$) in amounts represented by the coefficients b_p in (8.51). The effect of multiplying by $(\mathbf{A} - l^{(r)}\mathbf{I})$ or by $(\mathbf{A} - l^{(r)}\mathbf{I})^m$ can be described as 'purification' of the mixture by removal of most of the $\mathbf{x}^{(r)}$ component from it, and similarly for the effect of multiplication by $(\mathbf{A} - l^{(s)}\mathbf{I})^m$; this suggested the term 'purification process' used by Richardson.

In order that such a purification process should be effective, the values used for quantities like $l^{(r)}$, $l^{(s)}$ need only be approximations to characteristic values. This has two advantages in practical numerical work. First, for symmetrical matrices good approximations to characteristic values can be obtained, by use of formula (8.44), from approximations to characteristic vectors which are only moderate; hence if only one characteristic vector is required, it is not necessary to determine the others to any great accuracy. And, secondly, it is possible to use simple rounded values for the l 's if this would simplify the numerical work.

Example: To determine $\mathbf{x}^{(2)}$, $\mathbf{x}^{(3)}$ for the matrix

$$\mathbf{A} = \begin{bmatrix} -23 & 11 & 1 \\ 11 & -3 & -2 \\ 1 & -2 & 1 \end{bmatrix}$$

without accurate determination of $\mathbf{x}^{(1)}$.

Suppose that in the example of § 8.71 the successive approximation for $\mathbf{x}^{(1)}$, $\lambda^{(1)}$ has only been taken as far as the second stage, with $\mathbf{x}_{(1)} = (1, -0.48, -0.04)$, and it is desired to find the other characteristic vectors and values. The best available approximation to $\lambda^{(1)}$ at this stage is given by evaluating formula (8.44) for this vector $\mathbf{x}_{(1)}$, and is -28.00 . This, however, is so close to the actual value of $\lambda^{(1)}$ for

this matrix that use of it would not illustrate how effective the purification process can be although $l^{(1)}$ is only a rough approximation to $\lambda^{(1)}$. So in working the example here, the value $l^{(1)} = -27$ will be taken. Then

$$\mathbf{A} - l^{(1)}\mathbf{I} = \mathbf{A} + 27\mathbf{I} = \begin{bmatrix} 4 & 11 & 1 \\ 11 & 24 & -2 \\ 1 & -2 & 28 \end{bmatrix}.$$

This suggests $\mathbf{x} = (0, 1, 1)$ as a trial vector. For this vector \mathbf{x} ,

$$(\mathbf{A} + 27\mathbf{I})\mathbf{x} = (12, 22, 26),$$

for which the approximation $24(\frac{1}{2}, 1, 1)$ is adequate at this stage. Hence we start the approximation for the second characteristic vector with $\mathbf{x}_{(0)} = \frac{1}{2}(1, 2, 2)$. This gives

$$\mathbf{A}\mathbf{x}_{(0)} = \frac{1}{2}(1, 1, -1).$$

This differs considerably from $\mathbf{x}_{(0)}$, so that although this $\mathbf{x}_{(0)}$ is approximately orthogonal to $\mathbf{x}^{(1)}$, it is far from $\mathbf{x}^{(2)}$ (we shall see later that it is much more nearly in the direction of $\mathbf{x}^{(3)}$). It is also some way from being orthogonal to the best approximation available to $\mathbf{x}^{(1)}$, namely $(1, -0.48, -0.04)$, so that rather than using $(1, 1, -1)$ as the next trial vector $\mathbf{x}_{(1)}$, we repeat the purification process, which gives

$$(\mathbf{A} + 27\mathbf{I})(1, 1, -1) = (14, 37, -29) = 37(0.38, 1, -0.78)$$

and take $\mathbf{x}_{(1)} = (0.38, 1, -0.78)$ as the next trial vector. This gives

$$\mathbf{A}\mathbf{x}_{(1)} = (1.48, 2.74, -2.40) = 2.74(0.54, 1, -0.88) \quad (8.53)$$

and

$$(\mathbf{x}'_{(1)}\mathbf{A}\mathbf{x}_{(1)})/(\mathbf{x}'_{(1)}\mathbf{I}\mathbf{x}_{(1)}) = 2.95. \quad (8.54)$$

The further procedure depends on the results required. We will first consider the improvement of the approximation to $\mathbf{x}^{(2)}$, and then the determination of $\mathbf{x}^{(3)}$ without more information about $\mathbf{x}^{(2)}$ than is expressed by (8.53) and (8.54).

(a) *Improvement of the approximation to $\mathbf{x}^{(2)}$.* We now have the approximations $\lambda^{(1)} = -28.0$, $\lambda^{(2)} = +2.95$, and the general result that the sum of the characteristic values is the sum of the diagonal elements of the matrix, in this case -25 . Hence we can conclude that $\lambda^{(3)}$ is roughly 0.05 , so that $l^{(3)} = 0$ is a good approximation to $\lambda^{(3)}$, and multiplication by \mathbf{A} itself will be effective in removing from a trial vector \mathbf{x} any 'impurities' in the form of multiples of $\mathbf{x}^{(3)}$. Hence starting from $(0.54, 1, -0.88)$ given by (8.53) as the best approximation yet available to $\mathbf{x}^{(2)}$, we multiply by $(\mathbf{A} + 27\mathbf{I})$ to remove multiples of $\mathbf{x}^{(1)}$ and by \mathbf{A} to remove multiples of $\mathbf{x}^{(3)}$ and also to examine the successive approximations of $\mathbf{A}\mathbf{x}$ to $\lambda^{(2)}\mathbf{x}$.

Since $l^{(3)} = 0$ appears likely to be a better approximation to $\lambda^{(3)}$ than $l^{(1)} = 27$ is to $\lambda^{(1)}$, it will be best to carry out two or three multiplications by $(\mathbf{A} + 27\mathbf{I})$ for each multiplication by \mathbf{A} ; and since the vector $(0.54, 1, -0.88)$ is already the result of a multiplication by \mathbf{A} , and so contains only a small admixture of $\mathbf{x}^{(3)}$, let us start by some multiplications by $(\mathbf{A} + 27\mathbf{I})$:

$$\begin{aligned} (\mathbf{A} + 27\mathbf{I})(0.54, 1, -0.88) &= (12.28, 31.70, -26.10) \\ &= 31.70(0.387, 1, -0.823), \end{aligned}$$

$$(\mathbf{A} + 27\mathbf{I})(0.39, 1, -0.82) = 29.93(0.392, 1, -0.821),$$

$$(\mathbf{A} + 27\mathbf{I})(0.392, 1, -0.821) = 29.954(0.3922, 1, -0.8211).$$

This has produced a vector nearly free from $\mathbf{x}^{(1)}$ but possibly still containing traces of $\mathbf{x}^{(3)}$. To remove these we multiply by \mathbf{A} :

$$\mathbf{A}(0.3922, 1, -0.8211) = 2.9564(0.3918, 1, -0.8216).$$

The results of further multiplication by \mathbf{A} alone diverge. This can be seen as follows. Let \mathbf{x} be an approximation to $\mathbf{x}^{(2)}$:

$$\mathbf{x} = \mathbf{x}^{(2)} + b_1 \mathbf{x}^{(1)} + b_3 \mathbf{x}^{(3)},$$

where b_1 and b_3 are small compared with 1. Then

$$\mathbf{A}\mathbf{x} = \lambda^{(2)}[\mathbf{x}^{(2)} + (\lambda^{(1)}/\lambda^{(2)})b_1 \mathbf{x}^{(1)} + (\lambda^{(3)}/\lambda^{(2)})b_3 \mathbf{x}^{(3)}].$$

Now $|\lambda^{(1)}/\lambda^{(2)}|$ is greater than 1, and in this example it is about 10, so that any 'impurity' in \mathbf{x} in the form of a multiple of $\mathbf{x}^{(1)}$ is more prominent in $\mathbf{A}\mathbf{x}$ than in \mathbf{x} . This building up of $\mathbf{x}^{(1)}$ can be avoided by further multiplications by $(\mathbf{A} - l^{(1)}\mathbf{I})$:

$$(\mathbf{A} + 27\mathbf{I})(0.3918, 1, -0.8216) = 29.953(0.39215, 1, -0.82172),$$

$$(\mathbf{A} + 27\mathbf{I})(0.39215, 1, -0.82172) = 29.9571(0.39212, 1, -0.82171).$$

This vector $\mathbf{x} = (0.39212, 1, -0.82171)$ gives

$$\mathbf{A}\mathbf{x} = 2.9567(0.39217, 1, -0.82171),$$

$$(\mathbf{x}'\mathbf{A}\mathbf{x})/(\mathbf{x}'\mathbf{I}\mathbf{x}) = 2.9568.$$

So that to four decimals in $\lambda^{(2)}$ and a possible error of one or two in the fifth decimal in $\mathbf{x}^{(2)}$

$$\lambda^{(2)} = 2.9568, \quad \mathbf{x}^{(2)} = (0.39212, 1, -0.82171). \quad (8.55)$$

Notes: (i) The convergence would be much quicker if the value $l^{(1)} = -28$ instead of -27 were taken, but this working illustrates that the purification process is effective even if $l^{(1)}$ is only a rough approximation to $\lambda^{(1)}$.

(ii) With a matrix whose elements are integral, as in this example, the numerical work is simplified if the l 's are taken to have integral values; in this example all elements of \mathbf{A} and of $(\mathbf{A} - l^{(1)}\mathbf{I})$ are numbers of two digits only, which makes the multiplications very quick.

(b) *Determination of $\mathbf{x}^{(3)}$, $\lambda^{(3)}$.* The most striking illustration of the power of the method is provided by the determination of $\mathbf{x}^{(3)}$, the characteristic vector for the *smallest* value of $|\lambda|$, without requiring that $\mathbf{x}^{(1)}$ or $\mathbf{x}^{(2)}$ should be determined more accurately than they are at the stage of the calculation reached at formula (8.53) and (8.54); comparison of formula (8.53) with the result (8.55) shows that the approximation to $\mathbf{x}^{(2)}$ at that stage is decidedly rough.

We have used $l^{(1)} = -27$ as an approximation to $\lambda^{(1)}$, and can now adopt $l^{(2)} = +3$ as an approximation to $\lambda^{(2)}$; so repeated multiplication of an arbitrary vector \mathbf{x} by $(\mathbf{A} + 27\mathbf{I})(\mathbf{A} - 3\mathbf{I})$ will produce a sequence of vectors whose directions will converge to that of $\mathbf{x}^{(3)}$. Now

$$\begin{aligned} (\mathbf{A} + 27\mathbf{I})(\mathbf{A} - 3\mathbf{I}) &= \begin{bmatrix} 4 & 11 & 1 \\ 11 & 24 & -2 \\ 1 & -2 & 28 \end{bmatrix} \begin{bmatrix} -26 & 11 & 1 \\ 11 & -6 & -2 \\ 1 & -2 & -2 \end{bmatrix} \\ &= - \begin{bmatrix} -18 & 24 & 20 \\ 24 & 19 & 33 \\ 20 & 33 & 51 \end{bmatrix} = \mathbf{B} \quad (\text{say}), \end{aligned}$$

and we have seen that $\mathbf{x} = \frac{1}{2}(1, 2, 2)$ is nearly orthogonal to $\mathbf{x}^{(1)}$ and differs consider-

ably from $\mathbf{x}^{(2)}$, so let us take it as a first approximation to $\mathbf{x}^{(3)}$. Then successive multiplications by \mathbf{B} give:

$$\begin{aligned}\mathbf{B}(\tfrac{1}{2}, 1, 1) &= -(35, 64, 94) = -94(0.37, 0.68, 1), \\ \mathbf{B}(0.37, 0.68, 1) &= -(29.66, 54.80, 80.84) \\ &= -80.84(0.367, 0.678, 1), \\ \mathbf{B}(0.367, 0.678, 1) &= -80.711(0.3675, 0.6776, 1), \\ \mathbf{B}(0.3675, 0.6776, 1) &= -80.7108(0.36733, 0.67766, 1), \\ \mathbf{B}(0.36733, 0.67766, 1) &= -80.7094(0.36739, 0.67763, 1), \\ \mathbf{B}(0.36739, 0.67763, 1) &= -80.7096(0.36737, 0.67764, 1), \\ \mathbf{B}(0.36737, 0.67764, 1) &= -80.7095(0.36738, 0.67764, 1).\end{aligned}$$

This process could be continued indefinitely, to give as many decimals in $\mathbf{x}^{(3)}$ as might be required. With a possible error of 1 in the fifth decimal

$$\mathbf{x}^{(3)} = (0.36738, 0.67764, 1)$$

and with this value of $\mathbf{x}^{(3)}$

$$(\mathbf{x}^{(3)'}\mathbf{A}\mathbf{x}^{(3)})/(\mathbf{x}^{(3)'}\mathbf{I}\mathbf{x}^{(3)}) = 0.01209_3,$$

which is therefore an approximation to $\lambda^{(3)}$.

The collected results for this matrix \mathbf{A} are

$$\left. \begin{aligned}\lambda^{(1)} &= -27.969, & \mathbf{x}^{(1)} &= (1, -0.44579, -0.06530) \\ \lambda^{(2)} &= 2.9568, & \mathbf{x}^{(2)} &= (0.39212, 1, -0.82171) \\ \lambda^{(3)} &= 0.012093, & \mathbf{x}^{(3)} &= (0.36738, 0.67764, 1)\end{aligned}\right\}, \quad (8.56)$$

$\lambda^{(1)}, \mathbf{x}^{(1)}$ being given by (8.50) and $\lambda^{(2)}, \mathbf{x}^{(2)}$ by (8.55). The best check on these results is provided by verifying that the relations between characteristic vectors and between characteristic values are satisfied.

Since the characteristic vectors have been determined independently of one another, they can be checked by verifying that the orthogonality relations are satisfied. The results (8.56) give

$$\begin{aligned}\sum x_j^{(1)} x_j^{(2)} &= -0.00001, & \sum x_j^{(2)} x_j^{(3)} &= -0.00001, \\ \sum x_j^{(3)} x_j^{(1)} &= -0.00000_6,\end{aligned}$$

which differ from zero by amounts within the tolerance for rounding errors.

The sum of the characteristic values should be equal to the sum of the diagonal elements, which is -25 in this case, and this relation is satisfied exactly to three decimals. And since the elements of the matrix are integral, their determinant is integral, so the product of the characteristic values should be integral. For the results (8.56), $\lambda^{(1)}\lambda^{(2)}\lambda^{(3)} = -1.00007$, which differs from -1 by an amount within the tolerance for rounding errors.

A further check is provided by building up the matrix \mathbf{A} from its characteristic vectors and characteristic values according to formula (8.48); this, however, does not give a good check on $\mathbf{x}^{(3)}$, since its contribution is multiplied by the small factor $\lambda^{(3)}$.

Notes: (i) The ratio $|\lambda^{(1)}/\lambda^{(3)}|$ of the greatest and least in magnitude of the characteristic values is about 2300, and the large size of this quantity is an indication

of the degree to which the equations $\mathbf{Ax} = \mathbf{b}$, with this matrix \mathbf{A} , are ill-conditioned. This character of \mathbf{A} has not, however, introduced any difficulty in finding the characteristic values.

(ii) If the characteristic values and vectors are found in succession, the last (in this case the third) can be found as follows.

From an arbitrary vector \mathbf{x} , a vector \mathbf{X} , orthogonal to $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ is formed by taking

$$\mathbf{X} = (\mathbf{x} - \mu_1 \mathbf{x}^{(1)} - \mu_2 \mathbf{x}^{(2)}) \quad (8.57)$$

and determining μ_1, μ_2 appropriately. Since $\mathbf{x}^{(2)}$ is orthogonal to $\mathbf{x}^{(1)}$ it follows that the required values of μ_1, μ_2 are

$$\begin{aligned} \mu_1 &= \left(\sum_j x_j^{(1)} x_j \right) / \sum_j (x_j^{(1)})^2 \\ \mu_2 &= \left(\sum_j x_j^{(2)} x_j \right) / \sum_j (x_j^{(2)})^2 \end{aligned} \quad (8.58)$$

\mathbf{X} should then be multiplied by such a factor that its greatest component is 1, and the process repeated, as a check and to remove effects of rounding errors as far as possible.

If the elements of \mathbf{A} are integral, the following alternative process can sometimes be used to determine $\lambda^{(3)}$. As soon as $\lambda^{(1)}$ and $\lambda^{(2)}$ are found to the accuracy

$$\lambda^{(1)} = -27.969 \pm 0.001, \quad \lambda^{(2)} = 2.957 \pm 0.001,$$

it follows, from the relation

$$\lambda^{(1)} + \lambda^{(2)} + \lambda^{(3)} = \text{sum of diagonal elements} = -25,$$

that $\lambda^{(3)}$ lies in the range $\lambda^{(3)} = 0.012 \pm 0.002$,

and hence $\lambda^{(1)}\lambda^{(2)}\lambda^{(3)} = -0.99 \pm 0.17$.

But since the elements of \mathbf{A} are integral, this product must be integral, so must be -1 . Hence

$$\lambda^{(3)} = -1/\lambda^{(1)}\lambda^{(2)} = +0.01209.$$

(iii) The purification process can be used to hasten the convergence of the approximation to $\lambda^{(1)}$, $\mathbf{x}^{(1)}$. From the result (8.54) it follows that multiplication by $(\mathbf{A} - 3\mathbf{I})$ will be effective in removing multiples of $\mathbf{x}^{(2)}$ from a trial approximation to $\mathbf{x}^{(1)}$. Hence after the second step in the example as worked in § 8.71 we could proceed as follows:

$$(\mathbf{A} - 3\mathbf{I})(1, -0.442, -0.068) = -30.930(1, -0.44578, -0.06531),$$

$$\mathbf{A}(1, -0.44578, -0.06531) = -27.9689(1, -0.44578, -0.06530),$$

reaching the result in two fewer steps than in § 8.61.

This example illustrates some possibilities of Richardson's purification procedure; for developments of the idea, and other examples, reference should be made to Richardson's paper.†

† For some other methods, see C. Lanczos, *Journ. of Research, Nat. Bureau of Standards*, **45** (1950), 255; J. H. Wilkinson, *Proc. Camb. Phil. Soc.* **50** (1954), 536, and *Proceedings of a Symposium on Automatic Digital Computation*, N.P.L. 1953 (H.M.S.O. 1954), ch. 18; R. A. Brooker and F. H. Sumner, *Proc. Inst. Elect. Eng.* **103**, Part B (1956), Supplement No. 1, 114.

8.73. Relaxation process for characteristic vectors

Another method of determining characteristic vectors is a form of relaxation process.

If l is any given number and we try to find by a relaxation process a solution of the equations

$$\left. \begin{aligned} (a_{11}-l)x_1 + a_{12}x_2 + a_{13}x_3 + \dots &= 0 \\ a_{21}x_1 + (a_{22}-l)x_2 + a_{23}x_3 + \dots &= 0 \quad \text{etc.} \end{aligned} \right\} \quad (8.59)$$

in which not all the x_j 's are zero, then unless l is a characteristic value, it will not be possible to reduce all residuals to zero. But if the x_j 's are components of a characteristic vector, corresponding to a characteristic value λ , then

$$\begin{aligned} (a_{11}-l)x_1 + a_{12}x_2 + a_{13}x_3 + \dots &= (\lambda-l)x_1, \\ a_{21}x_1 + (a_{22}-l)x_2 + a_{23}x_3 + \dots &= (\lambda-l)x_2, \\ \cdot & \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \end{aligned}$$

so that the ratio [(residual of the j th equation)/ x_j] will be the same for all values of j .

If therefore we fix one of the x_j 's at a non-zero value, and in the relaxation process aim not at reducing all the residuals to zero but at making the residual of each j th equation proportional to x_j , then a change of l , forming an improvement in the approximation to λ , will have the effect of reducing all the residuals simultaneously.

If an approximate characteristic vector \mathbf{x} can be estimated on inspection of the equation, then an approximation Λ to the corresponding characteristic value is given by formula (8.44), and this can be taken as the best available value, at this stage of the calculation, to l .

Example: Consider again the matrix used in the examples in the previous sections, namely

$$\mathbf{A} = \begin{bmatrix} -23 & 11 & 1 \\ 11 & -3 & -2 \\ 1 & -2 & 1 \end{bmatrix}.$$

Inspection suggests that $\mathbf{x} = (2, -1, 0)$ is an approximation to a characteristic vector. For this \mathbf{x} , the vector \mathbf{Ax} is

$$\mathbf{Ax} = (-57, 25, 2)$$

and

$$\sum_{jk} x_j A_{jk} x_k / \sum_j (x_j)^2 = -139/5 = -27.8.$$

With $l = -27.8$ the matrix of the coefficients of equations (8.53) is

$$\mathbf{A} - l\mathbf{I} = \begin{bmatrix} 4.8 & 11 & 1 \\ 11 & 24.8 & -2 \\ 1 & -2 & 28.8 \end{bmatrix}$$

and the beginning of the relaxation process, keeping x_1 fixed, is as follows:

| | x_1 | x_2 | x_3 | R_1 | R_2 | R_3 | |
|------------|-------|-------|-------|-------|-------|--------|---------|
| Operations | 0 | 1 | 0 | 11 | 24.8 | -2 | |
| table | 0 | 0 | 1 | 1 | -2 | 28.8 | |
| <hr/> | | | | | | | |
| | x_1 | x_2 | x_3 | R_1 | R_2 | R_3 | |
| Relaxation | 2 | -1 | 0 | -1.4 | -2.8 | 4.0 | |
| table | | | -0.14 | -1.54 | -2.52 | -0.032 | |
| | | 0.1 | | -0.44 | -0.04 | -0.232 | |
| | | | 0.01 | -0.43 | -0.06 | 0.056 | |
| | | 0.01 | | -0.32 | 0.188 | 0.036 | |
| <hr/> | | | | | | | |
| | 2 | -0.89 | -0.13 | -0.32 | 0.188 | 0.036 | Checked |
| <hr/> | | | | | | | |

Although the residuals are not very closely proportional to the values of the x_j 's, they are now of such signs and magnitudes that a change of l which would reduce R_1 to zero would also decrease $|R_2|$ and $|R_3|$ considerably.

When a stage such as this has been reached, it is best to calculate a new approximation to λ ; in this case, with

$$\mathbf{x} = (2, -0.89, -0.13),$$

$$\mathbf{Ax} = (-55.92, 24.93, 3.65),$$

$$\sum_{jk} (x_j A_{jk} x_k) / \sum_j (x_j)^2 = 134.50/4.8090 = -27.968$$

and with this value of l , the residuals, to three decimals, are

$$0.016, \quad 0.039, \quad 0.014;$$

as expected, these values are substantially smaller than those at the end of the relaxation table above. A further relaxation can then be carried out starting from these values.

An alternative way of evaluating an improved approximation to λ is as follows. When residuals R_j have been obtained such that R_j is roughly proportional to x_j , then \mathbf{x} is an approximation to a characteristic vector of the matrix $(\mathbf{A}-l\mathbf{I})$, and a better approximation to the corresponding characteristic value $(\lambda-l)$ can be made by use of the formula (8.44) applied to the matrix $(\mathbf{A}-l\mathbf{I})$. Now the vector \mathbf{R} formed by the residuals R_j is given by $(\mathbf{A}-l\mathbf{I})\mathbf{x} = \mathbf{R}$, so an approximation to $\lambda-l$ is

$$\lambda-l = \sum_j x_j R_j / \sum_j (x_j)^2,$$

whence an improved approximation to λ is given by

$$\lambda = l + \sum_j x_j R_j / \sum_j (x_j^2). \tag{8.60}$$

A similar process can be carried out for the other characteristic vectors in order of decreasing $|\lambda|$. For these, however, it is necessary either to eliminate each one from the matrix as it is calculated, or to ensure that

each as it is calculated is made orthogonal to all those already determined, as in § 8.71.

Characteristic values of linear ordinary differential equations with two-point boundary conditions can be obtained by a combination of this technique with the replacement of derivatives by finite differences as in § 8.6.†

† See L. Fox, *Proc. Camb. Phil. Soc.* **45** (1948), 50, § 8.

IX

NON-LINEAR ALGEBRAIC EQUATIONS

9.1. Solution of algebraic equations

By an 'algebraic' equation is meant, in this chapter, an equation $f(x) = 0$ not involving derivatives or integrals of $f(x)$, and of which a solution is a number, as distinct from a *differential* equation of which a solution is a *function* of the continuous variable x . It does not imply that the function $f(x)$ whose zeros are to be found is an algebraic function. For example the equations

$$x^3 + 5x^2 - 3x - 2 = 0$$

and

$$e^x \sin x = 1$$

are both 'algebraic' in this sense.

$f(x)$ being a given function of x , the problem of finding the roots of $f(x) = 0$ is often best dealt with in two steps, the first concerned with locating the roots roughly, to two or three significant figures, and the second with improving these rough values.

The solution of an algebraic equation can be regarded as a process of inverse interpolation, for if $f(x)$ is tabulated as a function of x , then the determination of the value of x for which $f(x)$ has any given value, of which zero is a special case, is just the situation with which inverse interpolation is concerned. Once a solution has been located approximately, tabulation of the function in the neighbourhood of that solution, followed by a process of inverse interpolation, is one way of determining it more exactly. Another method is to use an iterative process; this is considered in § 9.3.

9.2. Graphical methods

Use of graphs is often a valuable method of locating approximately the roots of an equation $f(x) = 0$. Either the function $f(x)$ itself may be graphed and its intersections with the x -axis determined, or the equation may be written in the form $f_1(x) = f_2(x)$, and its roots determined by the intersection of the graphs of $y = f_1(x)$ and $y = f_2(x)$; it may be possible to avoid some calculation by this process. In some cases graphs using some argument other than x , such as $\log x$ or $1/x$, may be useful. The best procedure will depend on the form of the function $f(x)$, and it is difficult to lay down any general rules.

The following examples give suggestions for handling some kinds of equations:

(i) Equations of the form $f(x) \equiv x^n + ax + b = 0$. Use graphs of $y = x^n$ and $y = -(ax + b)$.

(ii) Equations of the form

$$f(x) \equiv x^n + a_1 x^{n-1} + a_2 x^{n-2} + \dots + a_{n-1} x + a_n = 0.$$

Evaluate $f(x)$ directly (see § 3.2) or by building up from its differences (see § 4.42) for, say, $x = -5(0.2) + 5$; for $|x| > 4$, say, take $y = 1/x$ and evaluate

$$F(y) = a_n y^n + a_{n-1} y^{n-1} + \dots + a_1 y + a_0.$$

Use graphs of $f(x)$ against x and $F(y)$ against y .

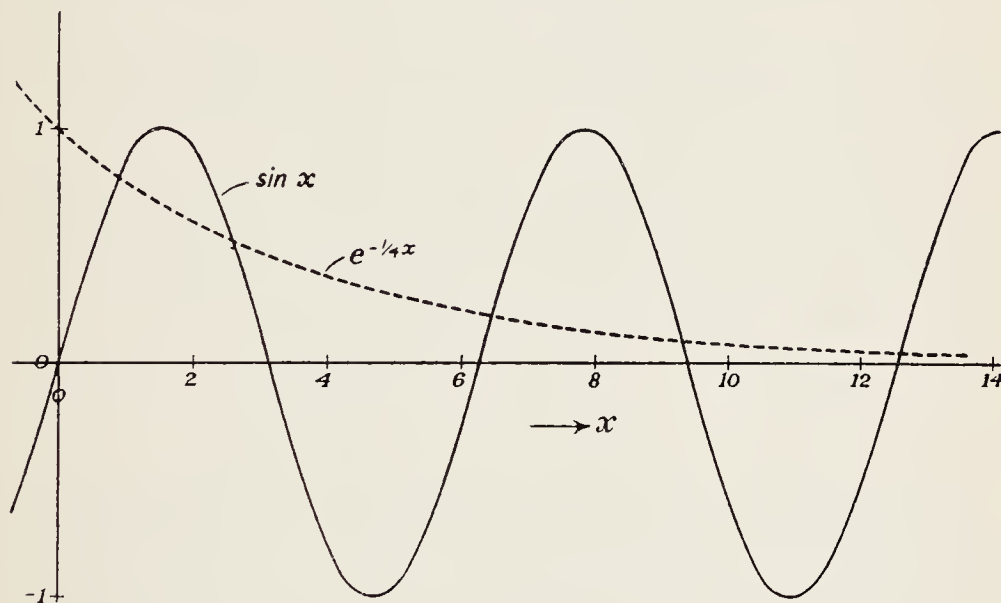


FIG. 13.

(iii) $e^{ix} \sin x = 1$. Write this as $\sin x = e^{-ix}$, and determine the intersections of $y = e^{-ix}$ with $y = \sin x$ (see Fig. 13). This avoids calculation of the products $e^{ix} \sin x$.

9.3. Iterative processes

By an iterative process is meant one in which the equation $f(x) = 0$ is expressed in the form

$$x = F(x),$$

and we try to find a solution by constructing a sequence $\{x_n\}$ by the relation

$$x_{n+1} = F(x_n). \quad (9.1)$$

If, to the degree of numerical accuracy to which the work is carried out, $x_{n+1} = x_n$, then this value of x_{n+1} is a solution of the equation to that degree of accuracy.

Let $x = X$ be the solution of the equation, and let

$$x_n = X + \xi_n \quad (9.2)$$

so that ξ_n is the error in x_n , regarded as a solution of the equation. An important feature of an iterative method is the way in which this error varies with the number n of repetitions of the iterative process. This can be examined by expanding the right-hand side of (9.1) in a Taylor series. Then, since $x = X$ satisfies $x = F(x)$, it follows that

$$\xi_{n+1} = a_1 \xi_n + a_2 \xi_n^2 + a_3 \xi_n^3 + \dots, \quad (9.3)$$

where $a_k = F^{(k)}(X)/k!$.

If $a_1 \neq 0$, then the errors ξ_n of results of successive repetitions of the iterative process are ultimately related by

$$\xi_{n+1} = a_1 \xi_n, \quad \xi_{n+m} = a_1^m \xi_n;$$

in order that the process should converge, $|a_1| = |F'(X)|$ must be less than 1, and the magnitude of the error then decreases exponentially with n increasing. This means that the number of additional correct significant figures obtained from each repetition of such a process (or, more often, the number of repetitions required to obtain each new correct significant figure) is the same, however many figures have been obtained. Such a process is called 'first-order'.

But if $a_1 = 0$, $a_2 \neq 0$ in (9.3), then the successive errors ξ_n are ultimately related by

$$\xi_{n+1} = a_2 \xi_n^2, \quad \text{whence} \quad a_2 \xi_{n+m} = (a_2 \xi_n)^{2^m},$$

where $a_2 = \frac{1}{2}F''(X)$. The number of correct significant figures is approximately doubled for each repetition of the iterative process, so that the better the approximation of x_n to X , the easier it is to improve it further. Such a process is called 'second-order', and once a fair approximation to $x = X$ has been attained, a second-order process is very greatly to be preferred to a first-order one; but it must be started from an approximation good enough to ensure that $|a_2 \xi_0| < 1$. It will be shown in § 9.32 that from any first-order process it is possible to derive a second-order process.

If $a_1 = 0$, $a_2 = 0$, $a_3 \neq 0$ in (9.3), then the successive errors ξ_n are ultimately related by

$$\xi_{n+1} = a_3 \xi_n^3, \quad \text{whence} \quad a_3 \xi_{n+m} = (a_3 \xi_n)^{3^m};$$

such a process is called 'third-order'. The formula for a third-order process is usually more complicated than that for a second-order process for the same equation, and the convergence of a second-order process is already so fast once a good approximation has been obtained that the advantage of still quicker convergence obtainable from a third-order process may be more than offset by the more complicated formulae

which have to be evaluated for each repetition of the iterative process, and third-order processes are not much used in practice. Second-order processes, however, are widely used.

9.31. Examples of iterative processes

(a) Newton's process for a square root

An important example of a second-order process is one for a square root, usually known as 'Newton's process'. If b is the number whose square root is required, this process consists of forming the sequence $\{x_n\}$ defined by

$$x_{n+1} = \frac{1}{2}[x_n + (b/x_n)].$$

For this process, $X = b^{\frac{1}{2}}$, and $F(x) = \frac{1}{2}[x + (b/x)]$, giving

$$F'(x) = \frac{1}{2}(1 - b/x^2), \quad F'(X) = 0,$$

$$F''(x) = b/x^3, \quad F''(X) = 1/X,$$

so that $a_1 = 0$, $a_2 \neq 0$ in (9.3), and the process is second order.

As an example of the application of this process, consider the evaluation of $\sqrt{12}$, starting from $x_0 = 2$.

$$x_0 = 2, \quad x_1 = \frac{1}{2}(2 + \frac{12}{2}) = 4,$$

$$x_1 = 4, \quad x_2 = \frac{1}{2}(4 + \frac{12}{4}) = \frac{7}{2} = 3.5,$$

$$x_2 = \frac{7}{2}, \quad x_3 = \frac{1}{2}(\frac{7}{2} + \frac{24}{7}) = \frac{97}{28} = 3.4643,$$

$$x_3 = 3.4643, \quad 12/x_3 = 3.46390,32, \quad x_4 = 3.46410,16,$$

$$x_4 = 3.46410,16, \quad 12/x_4 = 3.46410,16303, \quad x_5 = 3.46410,16151.$$

Notes: (i) For the first two or three iterations it may be easiest to work with the numbers in the form of rational fractions; later it is more convenient to work with them in decimal form.

(ii) In this example, an unnecessarily bad approximation has been taken as a starting-point to illustrate the convergence of the process from even a very poor value of x_0 . A more important application is the improvement of an already fairly good approximation such as x_3 . From Barlow's Tables a square root correct to four figures can always be obtained without any interpolation; then one application of Newton's process will give eight figures, and another will give fifteen figures at least.

Newton's process is not the only second-order one for a square root: another is given by

$$x_{n+1} = x_n(3b - x_n^2)/2b = x_n[1 + (b - x_n^2)/2b]. \quad (9.4)$$

This process does not converge as fast as Newton's, since for a given value of b , the value of a_2 in the series (9.3) is greater. But it has one feature which may be an advantage, namely that the divisor in (9.4) is constant instead of being different at each stage of the iteration process as it is in Newton's process.

(b) *The Newton-Raphson process*

A general second-order process for the solution of $f(x) = 0$, at a point not in the neighbourhood of a maximum or minimum of $f(x)$, is one given by

$$x_{n+1} = x_n - f(x_n)/f'(x_n). \quad (9.5)$$

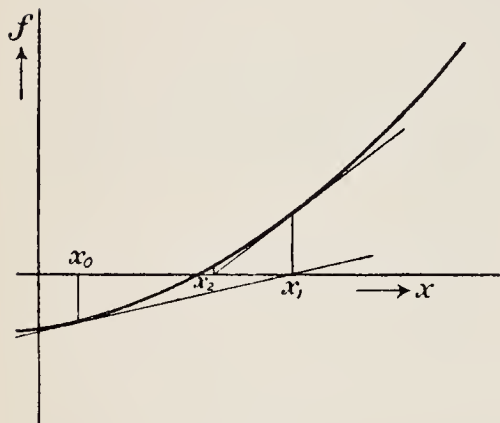


FIG. 14.

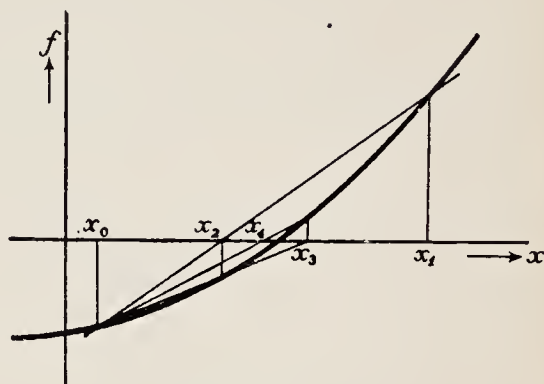


FIG. 15.

This is known as the Newton-Raphson process; Newton's process for a square root is the special case of it for the function $f(x) = x^2 - b$, and (9.4) is the special case of it for $f(x) = 1 - b/x^2$. For the general Newton-Raphson process,

$$F(x) = x - f(x)/f'(x), \quad F'(x) = -f(x)f''(x)/[f'(x)]^2,$$

and since $f(X) = 0$, it follows that $F'(X) = 0$, so the process is second order.

Expressed in terms of the graph of $f(x)$, the Newton-Raphson process is equivalent to linear interpolation along the tangent to the curve $y = f(x)$ at x_n (see Fig. 14).

(c) *The 'rule of false position'*

Another iterative process is equivalent to linear interpolation along the chord joining $[x_0, f(x_0)]$ to $[x_n, f(x_n)]$ (see Fig. 15). This gives

$$x_{n+1} = x_0 - (x_n - x_0)f_0/(f_n - f_0) = (x_0f_n - x_nf_0)/(f_n - f_0).$$

This method, however, is only first order, though if x_0 is a fair approximation to X , the coefficient a_1 in (9.3) is small; successive errors ξ_n are related by

$$\xi_{n+1} = \frac{1}{2}(x_0 - X)f''(X)\xi_n/f'(X),$$

approximately. It has the advantage that it does not require the evaluation of $f'(x)$.

A disadvantage of methods such as the Newton-Raphson and the method of false position is that they involve the evaluation of $f(x)$ and $f'(x)$ at a number of values of x which, though systematic in the sense that each is calculated from the previous one by the same formula such as (9.5), are irregularly spaced, and such a set of numbers is difficult to

check adequately. An advantage is that a mistake in an intermediate value of x_n does not affect the final result; it is just equivalent to starting a new iteration with this erroneous value of x_n as x_0 . But this does not eliminate the possibility of a mistake in the last repetition of the iterative process. Tabulation of $f(x)$ at equal intervals of x followed by a process of inverse interpolation is a process which provides more, and simpler, checks against occasional mistakes.

Example: To find the root of $x \tan x = \frac{1}{2}$ which lies between $x = 0.6$ and 0.7 .

(a) *By the Newton-Raphson process*

There are several forms in which this equation can be written, for example:

$$f(x) \equiv x \tan x - \frac{1}{2} = 0; \quad f(x) \equiv 2x - \cot x = 0;$$

$$f(x) \equiv 2x \sin x - \cos x = 0.$$

The third of these will be adopted, as it gives the most convenient formula for $f'(x)$, namely

$$f'(x) = 2x \cos x + 3 \sin x.$$

Starting with $x_0 = 0.6$, $\sin x_0 = 0.5646$, $\cos x_0 = 0.8253$, we have

$$\left. \begin{aligned} f(x_0) &= 2x_0 \sin x_0 - \cos x_0 = -0.1478 \\ f'(x_0) &= 2x_0 \cos x_0 + 3 \sin x_0 = +2.6842 \end{aligned} \right\}, \quad \begin{aligned} f(x_0)/f'(x_0) &= -0.0551, \\ x_1 &= 0.6 + 0.0551 \\ &= 0.6551, \end{aligned}$$

$$x_1 = 0.655, \sin x_1 = 0.609159, \cos x_1 = 0.793048$$

$$\left. \begin{aligned} f(x_1) &= 2x_1 \sin x_1 - \cos x_1 = +0.004950 \\ f'(x_1) &= 2x_1 \cos x_1 + 3 \sin x_1 = 2.86637 \end{aligned} \right\}, \quad \begin{aligned} f(x_1)/f'(x_1) &= +0.001727, \\ x_2 &= 0.655 - 0.001727 \\ &= 0.653273, \end{aligned}$$

$$x_2 = 0.653273, \sin x_2 = 0.607788, \cos x_2 = 0.794099$$

$$\left. \begin{aligned} f(x_2) &= 2x_2 \sin x_2 - \cos x_2 = +0.000007_7 \\ f'(x_2) &= 2x_2 \cos x_2 + 3 \sin x_2 = 2.86097 \end{aligned} \right\}, \quad \begin{aligned} f(x_2)/f'(x_2) &= +0.000002_7, \\ x_3 &= 0.653270. \end{aligned}$$

Notes: (i) The first approximation $x_0 = 0.6$ is a rough one and four-figure values of $\sin x$, $\cos x$ are adequate at this stage; more figures are used later when the accuracy of x_n has been improved.

(ii) For the second stage of the iteration, x_1 is taken as 0.655 instead of the value 0.6551 obtained from the first stage. It is not to be expected that the fourth decimal of this value will be correct, and the rounded value $x_1 = 0.655$ enables tables with interval $\delta x = 0.001$ † to be used without interpolation. For the third stage, however, interpolation in the tables is necessary.

(iii) For the third stage (and later stages, if any) it would be adequate to use $f'(x_1)$ instead of recalculating $f'(x_n)$ for each new value of x_n . This makes the method formally only first order, but the coefficient a_1 in (9.3) is so small in such a case that the convergence of the first-order process is adequate for practical work.

(b) *By inverse interpolation*

To solve the equation by inverse interpolation, it is most convenient to take it in the form

$$f(x) \equiv 2x - \cot x = 0$$

† *Chambers's 6-Figure Tables*, vol. 2 (1949), for example.

as this avoids the calculation of any products and involves the least reference to tables. Evaluation of $f(x)$ to two or three decimals at intervals 0.05 or 0.02 in the range $x = 0.6$ to 0.7 , as might be used for a rough plot, locates the root as lying between $x = 0.65$ and 0.66 , and taking 0.01 intervals we have the following table:

| | $\cot x$ | $f(x) = 2x - \cot x$ | $\delta^2 f$ |
|------|----------|----------------------|--------------|
| 0.64 | 1.343104 | -0.063104 | |
| | | | 47668 |
| 0.65 | 1.315436 | -0.015436 | -718 |
| | | | 46950 |
| 0.66 | 1.288486 | +0.031514 | -685 |
| | | | 46265 |
| 0.67 | 1.262221 | +0.077779 | |

Inverse interpolation for $f(x) = 0$ at $x = 0.65 + p(0.01)$ gives

$$p = \frac{1.5436}{4.6950} - B^{(II)}(p) \cdot \frac{1.403}{4.6950}$$
$$= 0.32878 - B^{(II)}(p) \cdot (0.02988),$$

and iterative solution of this equation gives $p = 0.32715$, $x = 0.653271_5$.

Alternatively, 0.002 or 0.001 intervals could be taken between $x = 0.65$ and 0.66 ; for example

| | $\cot x$ | $f(x) = 2x - \cot x$ | |
|-------|----------|----------------------|------|
| 0.650 | 1.315436 | -0.015436 | |
| | | | 9447 |
| 0.652 | 1.309989 | -0.005989 | -29 |
| | | | 9418 |
| 0.654 | 1.304571 | +0.003429 | -29 |
| | | | 9389 |
| 0.656 | 1.299182 | +0.012818 | |

and linear interpolation is now adequate to give five decimals in x .

Note: The two methods may be combined; for example after obtaining the approximation x_1 by the Newton-Raphson method, the approximation to the root may be improved by tabulation at 0.001 intervals in the neighbourhood of $x = 0.655$, followed by inverse interpolation. This avoids the interpolation in tables which has to be done if the Newton-Raphson process is continued.

9.32. Derivation of a second-order process from a first-order process

If it is known that an iterative process is first order, this knowledge enables a better approximation to the solution to be obtained by an application of the process of ‘exponential extrapolation’ (see § 3.4 (a)). If the first term in the expansion (9.3) were the only term, then we would have

$$\xi_2/\xi_1 = \xi_1/\xi_0 = a_1, \quad \text{that is, } \xi_0 \xi_2 = \xi_1^2$$

exactly, so that X would be given by

$$(x_2 - X)(x_0 - X) = (x_1 - X)^2.$$

Unless the higher terms in (9.3) are negligible, this will not give exactly the value of X , but an approximation, say X^* , to it:

$$X^* = \frac{x_2 x_0 - x_1^2}{x_2 - 2x_1 + x_0} = x_2 - \frac{(x_2 - x_1)^2}{x_2 - 2x_1 + x_0} \tag{9.6}$$

(see eq. 3.12), which will usually be a substantially better approximation than x_2 . We can then repeat the process starting with $x_0 = X^*$.

In general, let X_{n+1}^* be the result of this process, starting with $x_0 = X_n^*$. Then it can be shown that the process of forming the successive values of X^* is second order.†

Example: To solve the equation $x^2 - 6x + 2 = 0$ by writing it in the form

$$x = 6 - (2/x)$$

and using an iterative process.

In this case the function $F(x)$ of § 9.3 is $F(x) = 6 - (2/x)$, $F'(x) = 2/x^2$ and though the solution $x = X$ is not yet known, we can be sure that $F'(x)$ is not zero there, hence the iterative process $x_{n+1} = 6 - (2/x_n)$ is first order.

Starting with $X_0^* = 3$ we have

$$x_0 = 3, \quad x_1 = \frac{16}{3}, \quad x_2 = \frac{45}{8}.$$

Then, from formula (9.6),

$$\begin{aligned} x_2 - x_1 &= \frac{7}{24}, & x_2 - 2x_1 + x_0 &= -\frac{49}{24}, \\ X_1^* &= \frac{45}{8} - \frac{(7/24)^2}{-49/24} = \frac{17}{3} = 5.66667. \end{aligned}$$

Then with $x_0 = X_1^* = \frac{17}{3}$, we have

$$x_1 = 6 - 2.3/17 = 5.64706, \quad x_2 = 6 - 2/5.64706 = 5.64584;$$

then from formula (9.6) again,

$$\begin{aligned} x_2 - x_1 &= -0.00122; & x_2 - 2x_1 + x_0 &= 0.01839; \\ X_2^* &= 5.64584 - \frac{(0.00122)^2}{0.01839} = 5.64584 - 0.00008 = 5.64576, \end{aligned}$$

which is only in error by 1 in the fifth decimal place.

9.4. Multiple roots and neighbouring roots

Particular care is necessary when the coefficients in the equation are in the neighbourhood of values for which the equation has multiple roots. The values of the roots are then particularly sensitive to the values of the coefficients and to rounding errors. To take a simple example, the equation

$$x^2 - 8x + 16.00 = 0 \text{ has a double root } x = 4,$$

$$x^2 - 8x + 16.01 = 0 \text{ has no real root,}$$

$$x^2 - 8x + 15.99 = 0 \text{ has roots } 3.9, 4.1,$$

so that a change of less than one part in 1000 in the constant term affects the roots by one part in 40. The situation is clear here, but may not be when the equation has other roots or involves transcendental functions. If a repeated root is suspected, either from the results of this process or from the graph, then careful numerical evaluation of the function should be carried out in the neighbourhood of the suspected repeated root.

† See D. R. Hartree, *Proc. Camb. Phil. Soc.* **45** (1948), 230.

Two (or more) close but not equal roots may be more troublesome than a true repeated root. If two close roots are suspected from examination of a graph or on other evidence, there will certainly be a root of $f'(x)$ in this neighbourhood, say $x = x_m$, and it is best to determine this first, and to evaluate $f(x_m)$. If this has the same sign as $f(x)$ at neighbouring values of x , then there is no real root; if $f(x_m)$ has the opposite sign to $f(x)$ at neighbouring values of x , then there are two real roots. Since $f'(x_m) = 0$, Taylor's series for $f(x)$ in the neighbourhood of x_m begins

$$f(x) = f(x_m) + \frac{1}{2}(x - x_m)^2 f''(x_m),$$

and if $(x - x_m)$ is sufficiently small (and $f''(x_m)$ is not too small)

$$x - x_m = \pm [2\{f(x) - f(x_m)\}/f''(x_m)]^{\frac{1}{2}}. \quad (9.7)$$

This, if not already accurate enough, will provide starting values for a further approximation to these roots.

The calculated values of these roots will be very sensitive to rounding errors; if $f(x)$ is given by a formula which can be evaluated to any degree of numerical accuracy (for example, by a polynomial), the roots can be evaluated to any accuracy required; but if the evaluation of $f(x)$ involves reference to tables, the accuracy of the calculated values of the roots may be small.

Example: To find the smallest positive root of $x \cos^2 x = 0.4115$.

A graph shows that $f(x) = x \cos^2 x$ has a maximum of about 0.41 in the neighbourhood of $x = 0.65$.

For this function, $f'(x) = -\cos x(2x \sin x - \cos x)$, and this has a zero at $x_m = 0.65327$ (see § 9.31, example), where $f(x) = 0.411949$; also

$$f''(x_m) = -\cos x_m(2x_m \cos x_m + 3 \sin x_m) = -2.272.$$

So $f(x) - f(x_m) = 0.4115 - 0.411949 = -0.000449$,

and the smallest root of $x \cos^2 x = 0.4115$ is approximately

$$\begin{aligned} 0.65327 - [2(-0.000449)/(-2.272)]^{\frac{1}{2}} &= 0.65327 - 0.01988 \\ &= 0.63339. \end{aligned}$$

Note: The fifth decimal of this value of x is not determined to ± 1 by the sixth decimal of $f(x)$ or $f(x_m)$.

9.5. Special processes for special types of equations

The methods so far considered have been general methods applicable to any kind of algebraic equation, in the sense explained in § 9.1. For some special kinds of equations $f(x) = 0$, and particularly for those in which $f(x)$ is a polynomial in x , there are special methods.

Polynomial equations, like linear simultaneous equations, are met in contexts of two kinds. In one the coefficients are all known exactly and are usually integral; in the other they are only known to within a certain

tolerance because they are results either of observations subject to experimental errors or of other calculations which are subject to rounding errors.

In the latter case it may be important to know the range of uncertainty of the solution arising from the tolerance in one or more of the coefficients. If the polynomial is

$$f(x) \equiv a_0 x^n + a_1 x^{n-1} + a_2 x^{n-2} + \dots + a_{n-1} x + a_n = 0 \quad (9.8)$$

consider the first-order variation of a root $x = X$ with one of the coefficients, a_k . If the root X changes by ΔX when a_k changes by Δa_k , then

$$f'(X) \Delta X + X^{n-k} \Delta a_k = 0,$$

that is,

$$\Delta X = -[X^{n-k}/f'(X)] \Delta a_k,$$

and for changes (not necessarily equal) in all the coefficients

$$\Delta X = -\left[\sum_k (X^{n-k} \Delta a_k) \right] / f'(X). \quad (9.9)$$

This shows that the roots are particularly sensitive to the values of the coefficients in the neighbourhood of a stationary value of the function.

9.51. Quadratic equations

The roots of a quadratic equation $ax^2 + bx + c = 0$ can be evaluated from the standard formula

$$x = [-b \pm (b^2 - 4ac)^{1/2}] / 2a, \quad (9.10)$$

and this is probably the best way of evaluating complex roots. However, as already pointed out (§ 3.4 (b)), this is *not* always the best way of determining numerical values of the roots when they are real, and particularly not if the ratio of the roots is large ($b^2 \gg 4ac$).

A better practical method in many cases is an iterative process based on use of the relations

$$x_1 + x_2 = -b/a, \quad x_1 x_2 = c/a,$$

where x_1 and x_2 are the two roots. If x_1 is the root of greater modulus, then successive approximations to the roots can be evaluated by using the formulae

$$x_1 = -(b/a) - x_2, \quad x_2 = (c/a)/x_1 \quad (9.11)$$

alternately, starting from the approximation $x_2 = 0$ if no other is easily available.

This process, though only first order, can be carried out so easily, and when $b^2 \gg 4ac$ it converges so quickly, that it is unnecessary to refine it further. If b^2 is not considerably larger than $4ac$, it may be convenient to use a second-order process derived from this first-order process as

explained in § 9.32. Elimination of x_2 between the two equations (9.11) gives

$$x_1 = -(b/a) - (c/a)/x_1,$$

which is the general expression of which the example given in § 9.32 is a special case.

9.52. Cubic and quartic equations

A cubic equation with real coefficients has at least one real root, say x_1 ; if it is determined, division of the cubic by $(x - x_1)$ gives a quadratic which can then be solved by the standard formula or by iteration. For determination of the real root or roots the general methods of the previous sections will often be best.

There are also special methods available which depend on reducing the cubic to a standard form. If the cubic is

$$ax^3 + bx^2 + cx + d = 0$$

then the substitution $y = \beta(x + b/3a)$ reduces it to a cubic in y without a y^2 term; then either the coefficient of the term in y is zero, or a real value of β can be chosen so that the ratio of the coefficients of the terms in y and in y^3 is either $+1$ or -1 . Thus any cubic can be reduced to one of the forms

$$y^3 + D = 0, \quad y^3 + y + D = 0, \quad y^3 - y + D = 0;$$

the solution to an equation of the first of these forms can be found directly from a table of cube roots, and tables of the roots of the equations of the other two forms have been evaluated (for details, see the *Index of Mathematical Tables*).

If in the substitution $y = \beta(x + b/3a)$, β is chosen so as to reduce the equation to the form

$$4y^3 \pm 3y - D = 0$$

the further substitution $y = \sinh u$ (if the sign of the middle term is $+$), $y = \cosh u$ (if the sign of the middle term is $-$, and $D > 1$) or $y = \cos u$ (if the sign of the middle term is $-$, and $D < 1$) can be used to reduce it further to

$$\sinh 3u = D, \quad \cosh 3u = D, \quad \text{or} \quad \cos 3u = D$$

respectively and the solution found from tables of hyperbolic or circular functions.

For a quartic equation

$$ax^4 + bx^3 + cx^2 + dx + e = 0,$$

the cubic term can be removed by the substitution $y = (x + b/4a)$. Let the resulting equation be

$$y^4 + Cy^2 + Dy + E = 0$$

and let $(y^2 - \alpha y + \beta)$, $(y^2 + \alpha y + \gamma)$ be quadratic factors of the left-hand side. Then, multiplying these factors and equating coefficients,

$$\beta + \gamma - \alpha^2 = C, \tag{9.12}$$

$$\alpha(\beta - \gamma) = D, \tag{9.13}$$

$$\beta\gamma = E. \tag{9.14}$$

Elimination of β and γ gives a cubic equation for α , but instead of carrying out this elimination algebraically, it is more convenient to proceed as follows. From equations (9.12), (9.13)

$$\beta + \gamma = C + \alpha^2, \quad \beta - \gamma = D/\alpha. \tag{9.15}$$

From these, β and γ can be evaluated for a set of values of α , and the value of α

for which $4\beta\gamma = 4E$ in agreement with (9.14) can be determined by trial and inverse interpolation. If the roots of the quartic are not all real, this value of α is unique; if the roots are all real, there are three values of α corresponding to the three ways in which the real linear factors of the left-hand side can be separated into two pairs.

Example: Find the roots of the equation

$$x^4 - 6.4x^3 + 19.8x^2 - 31.5x + 25 = 0.$$

The substitution of $y = x - 1.6$ gives the equation for y

$$y^4 + 4.44y^2 - 0.908y + 5.6272 = 0,$$

so equations (9.15), (9.14) become

$$\beta + \gamma = 4.44 + \alpha^2, \quad \beta - \gamma = -0.908/\alpha, \quad 4\beta\gamma = 22.5088.$$

A first trial set of values of α gives

| α | $\beta + \gamma$ | $\beta - \gamma$ | 2β | 2γ | $4\beta\gamma$ |
|----------|------------------|------------------|----------|-----------|----------------|
| 0.2 | 4.48 | -4.54 | -0.06 | +9.02 | -0.54 |
| 0.3 | 4.53 | -3.027 | +1.503 | 7.557 | +11.36 |
| 0.4 | 4.60 | -2.27 | 2.33 | 6.87 | 16.01 |
| 0.5 | 4.69 | -1.816 | 2.874 | 6.506 | 18.70 |
| 0.6 | 4.80 | -1.513 | 3.287 | 6.313 | 20.75 |
| 0.7 | 4.93 | -1.297 | 3.633 | 6.227 | 22.62 |
| 0.8 | 5.08 | -1.135 | 3.945 | 6.215 | 24.52 |

These values indicate that to give $4\beta\gamma = 22.5088$, the value of α is about 0.694, and further trial values in this neighbourhood give

| | | | | | | |
|-------|--------|---------|--------|--------|--------|----|
| 0.693 | 4.9202 | -1.3102 | 3.6100 | 6.2304 | 22.492 | 18 |
| 0.694 | 4.9216 | -1.3084 | 3.6132 | 6.2300 | 22.510 | |
| 0.695 | 4.9230 | -1.3065 | 3.6165 | 6.2295 | 22.529 | 19 |

from which $\alpha = 0.6939_5$; this value can be improved if required by taking the approximation farther. The evaluation of the roots of the equation for y (which in this example are all complex) from formula (9.10) is straightforward, and from these the roots of the original equation for x follow directly.

9.53. Polynomial equations

For solving polynomial equations there are available a number of special methods.† Various theorems in the theory of equations can be used to determine how many roots lie in various ranges of x , and various special methods are available such as Horner's method or a method known as 'root-squaring'‡ which depends on forming an equation whose roots are some high power of the roots of the equation to be solved.

The root-squaring method, however, only gives the magnitudes of the roots and not their signs, and some evaluation of $f(x)$ is necessary in order to determine their signs. Also both methods, as usually presented,

† For a survey, with particular reference to equations of high degree, see F. W. J. Olver, *Phil. Trans. Roy. Soc. A*, **244** (1952), 385.

‡ See E. T. Whittaker and G. Robinson, *Calculus of observations* (Blackie, 4th ed., 1944), § 64.

are deficient or lacking in current checks, and their results should always be verified either by substitution in $f(x)$ or by evaluating $f(x)$ at a numerically convenient set of values in the neighbourhood of each root followed by interpolation. The root-squaring method in particular offers too many opportunities for mistakes for any alleged root to be accepted without such investigation.

Some evaluation of $f(x)$ is therefore required in any case, and it seems better for most practical purposes to use the general methods already considered, namely the use of graphs for approximate location of the roots followed either by evaluation in the neighbourhood of each root and inverse interpolation, or by an iterative process,† rather than to use the special methods available for polynomial equations.

9.54. Repeated roots

Repeated roots of polynomial equations can be located by finding the highest common factor of $f(x)$ and $f'(x)$. If the coefficients in the equations are known exactly (when they will usually be integral), repeated roots can be identified with certainty if the H.C.F. process is carried out exactly without any rounding off. Otherwise there can generally be no certainty whether the equation has a repeated root or two (or more) very nearly equal roots. However the H.C.F. process may well establish the absence of a repeated root.

With polynomial equations repeated roots should be removed by dividing $f(x)$ by the appropriate product of repeated factors before the determination of the remaining roots is started.

9.55. Division of a polynomial by a quadratic

A convenient way of finding the complex roots of a real polynomial equation with no real roots is to express the polynomial as a product of real quadratic factors. In one process for doing this it is necessary to carry out a number of divisions by successive approximations to a real quadratic factor, and it will be convenient first to consider a numerical process for carrying out this division.

Let $a_0 x^n + a_1 x^{n-1} + a_2 x^{n-2} + \dots + a_{n-1} x + a_n$

be the dividend polynomial, and

$$x^2 + d_1 x + d_2$$

the divisor quadratic. Then we want to find a quotient polynomial

$$q_0 x^{n-2} + q_1 x^{n-3} + q_2 x^{n-4} + \dots + q_{n-3} x + q_{n-2}$$

† For an iterative process for complex roots of polynomial equations, see P. A. Samuelson, *Journ. Math. and Phys.* 28 (1949), 259.

Example: To divide $x^4 + 5x^3 + 12x^2 + 14x + 8$ by $x^2 + 2x + 4$ and by $1 + \frac{3}{2}x + \frac{1}{2}x^2$.

| Coefficient of | | | | | | | | | | | |
|----------------------------|-------|-------|-------|--------------------|----|---|----------------|-----------------------------|-----|---|----------------|
| | x^4 | x^3 | x^2 | | | x^4 | x^3 | x^2 | x | 1 | |
| 1 | 1 | 5 | 12 | 14 | 8 | 1 | 5 | 12 | 14 | 8 | 1 |
| -2 | | -2 | -6 | -4 | | | $-\frac{1}{2}$ | -3 | -12 | | $-\frac{3}{2}$ |
| -4 | | | 4 | 12 | -8 | $-\frac{5}{2}$ | -1 | -4 | | | $-\frac{1}{2}$ |
| Sum | 1 | 3 | 2 | -2 | 0 | $-\frac{3}{2}$ | $-\frac{7}{2}$ | 5 | 2 | 8 | Sum |
| Quotient $x^2 + 3x + 2$ | | | | Remainder $-2x$ | | Remainder $-\frac{7}{2}x^3 - \frac{3}{2}x^4$ | | Quotient $8 + 2x + 5x^2$ | | | |

9.56. Real quadratic factors of a polynomial

If some of the roots of a real polynomial equation $f(x) = 0$ are complex, it is best first to determine the real roots x_1, x_2, \dots, x_k by the methods already considered, and remove them by dividing $f(x)$ by $(x - x_1)(x - x_2) \dots (x - x_k)$; repeated roots, if any, should be identified as explained in § 9.54, and taken out with the corresponding multiplicity. This process provides a good check on these roots, since if the remainder in the division differs from zero by a greater amount than the tolerance for rounding errors, one of the roots must be in error.

The quotient will be a polynomial $F(x)$ with no real zeros, and its complex zeros can be determined by finding its real quadratic factors. We take a trial quadratic $D_0 = x^2 + b_0x + c_0$ and find the quotient Q_0 on forward division of $F(x)$ by D_0 (the remainder is irrelevant). Then we find the quotient Q_0^* on backward division of $F(x)$ by Q_0 (or a multiple of it), write

$$D_1 = Q_0^*/(\text{coeff. of } x^2 \text{ in } Q_0^*) = x^2 + b_1x + c_1,$$

and repeat the process with D_1 in place of D_0 . This provides an iterative process† in which the successive quadratics

$$D_0 = x^2 + b_0x + c_0,$$

$$D_1 = x^2 + b_1x + c_1,$$

$$D_2 = x^2 + b_2x + c_2$$

converge to the quadratic factor corresponding to the roots of smallest modulus of the equation $F(x) = 0$. The condition for convergence is that these roots should be of smaller modulus than any other roots of the equation; one of the reasons for removing the real roots is to ensure that there shall be no real root of smaller modulus than any of the complex

† B. Friedman, *Commun. on Pure and Appl. Math.* **2** (1949), 195. For another iterative process see A. C. Aitken, *Proc. Roy. Soc. Edin.* **63** (1951), 174.

roots, as the presence of such a root would make the results of this iterative process oscillate instead of converging.

Example: To find the real quadratic factors of

$$x^4 + 5x^3 + 12x^2 + 14x + 8.$$

For a quartic the method of § 9.52 using equations (9.14) and (9.15) would be simpler than the method of this section. The use of a quartic in this example is for illustration only.

The first steps, starting from $x^2 + 2x + 4$ as trial quadratic factor, have already been carried out in the example of the last section. The quotient of the first forward division is $x^2 + 3x + 2 = 2(1 + \frac{3}{2}x + \frac{1}{2}x^2)$; the factor 2 is taken out in order that the constant term in this divisor should be unity, and $1 + \frac{3}{2}x + \frac{1}{2}x^2$ used as the divisor in the backward division. For the present purpose the remainders are irrelevant, and the two pairs of columns in the centre can be omitted, so that thus far the working could be written

| Coefficient of | x^4 | x^3 | x^2 | x^2 | x | 1 | |
|----------------|--|-------|-------|------------------------------|-----|---|----------------|
| 1 | 1 | 5 | 12 | 12 | 14 | 8 | 1 |
| -2 | | -2 | -6 | -3 | -12 | | $-\frac{3}{2}$ |
| -4 | | | -4 | -4 | | | $-\frac{1}{2}$ |
| <hr/> | | | | | | | |
| | 1 | 3 | 2 | 5 | 2 | 8 | |
| | $= 2(\frac{1}{2} \quad \frac{3}{2} \quad 1)$ | | | $= 5(1 \quad 0.4 \quad 1.6)$ | | | |

The calculation proceeds as follows:—

| Coefficient of | x^4 | x^3 | x^2 | x^2 | x | 1 | |
|----------------|---------------------------------|-------|-------|--------------------------------|-----|---|------|
| 1 | 1 | 5 | 12 | 12 | 14 | 8 | 1 |
| -0.4 | | -0.4 | -1.84 | -5 | -4 | | -0.5 |
| -1.6 | | | -1.6 | -0.8 | | | -0.1 |
| <hr/> | | | | | | | |
| | 1 | 4.6 | 8.56 | 6.2 | 10 | 8 | |
| | $= 8.56(0.1 \quad 0.5 \quad 1)$ | | | $= 6.2(1 \quad 1.6 \quad 1.3)$ | | | |

| Coefficient of | x^4 | x^3 | x^2 | x^2 | x | 1 | |
|----------------|-----------------------------------|-------|-------|-----------------------------------|-------|---|-------|
| 1 | 1 | 5 | 12 | 12 | 14 | 8 | 1 |
| -1.6 | | -1.6 | -5.44 | -5.72 | -5.20 | | -0.65 |
| -1.3 | | | -1.3 | -1.52 | | | -0.19 |
| <hr/> | | | | | | | |
| | 1 | 3.4 | 5.26 | 4.76 | 8.80 | 8 | |
| | $= 5.26(0.19 \quad 0.65 \quad 1)$ | | | $= 4.76(1 \quad 1.85 \quad 1.68)$ | | | |

| Coefficient of | x^4 | x^3 | x^2 | x^2 | x | 1 | |
|----------------|-------------------------------------|-------|-------|-------------------------------------|-------|---|--------|
| 1 | 1 | 5 | 12 | 12 | 14 | 8 | 1 |
| -1.85 | | -1.85 | -5.83 | -5.88 | -5.62 | | -0.702 |
| -1.68 | | | -1.68 | -1.78 | | | -0.223 |
| <hr/> | | | | | | | |
| | 1 | 3.15 | 4.49 | 4.34 | 8.38 | 8 | |
| | $= 4.49(0.223 \quad 0.702 \quad 1)$ | | | $= 4.34(1 \quad 1.931 \quad 1.843)$ | | | |

The last three approximations to a quadratic factor are

$$x^2 + 1.6x + 1.3,$$

$$x^2 + 1.85x + 1.68,$$

$$x^2 + 1.931x + 1.843.$$

The process is first order, so we can use the method of 'exponential extrapolation' (see §§ 3.4 (a) and 9.32) to estimate a better approximation from these. If we write the quadratic x^2+bx+c , with suffixes 0, 1, 2 for these approximations, and use formula (9.6), we have

$$b_2-b_1 = 0.081, \quad b_2-2b_1+b_0 = -0.169,$$

$$\text{extrapolated } b = 1.931 + \frac{(0.081)^2}{0.169} = 1.970;$$

$$c_2-c_1 = 0.163, \quad c_2-2c_1+c_0 = -0.217,$$

$$\text{extrapolated } c = 1.843 + \frac{(0.163)^2}{0.217} = 1.965;$$

and the calculation can be continued from these values in a similar way. An alternative method of improving the approximation to a real quadratic factor will be considered in the following section.

Notes: (i) As illustrated in the working of this example, only a few significant figures need be kept at first when the approximation to a quadratic factor is still only rough, and more kept as the calculation proceeds.

(ii) When applied to a $2n$ th degree polynomial, the quotient of the forward division is a polynomial of the $(2n-2)$ th degree and in the backward division we have to divide by this quotient. But the quotient of the backward division is a quadratic and is determined by the leading three terms in the divisor in this backward division, so that the above process for division by a quadratic can still be used.

(iii) There is no accumulation of rounding errors, since at each stage the *original polynomial* is divided by the current trial quadratic factor.

9.57. Second-order process for improving the approximation to a quadratic factor

The following is an extension of the Newton-Raphson process to the improvement of an approximation to a real quadratic factor of a real polynomial $f(x)$.

Let (x^2+bx+c) be an approximate quadratic factor and let

$$f(x) = (x^2+bx+c)q(x) + rx + s, \quad (9.16)$$

where $q(x)$ is the quotient polynomial on division of $f(x)$ by (x^2+bx+c) , and $(rx+s)$ is the remainder. These can be found by the method of § 9.55. Differentiation of (9.16) with respect to b , for constant x and c , gives the variation of the coefficients r and s in the remainder with variation of the coefficient b in the trial quadratic factor:

$$0 = (x^2+bx+c)\left(\frac{\partial q(x)}{\partial b}\right) + xq(x) + \left(\frac{\partial r}{\partial b}\right)x + \left(\frac{\partial s}{\partial b}\right),$$

so that $-(\partial r/\partial b)x - (\partial s/\partial b)$ is the remainder when $xq(x)$ is divided by (x^2+bx+c) . Similarly differentiation with respect to c gives

$$0 = (x^2+bx+c)\left(\frac{\partial q(x)}{\partial c}\right) + q(x) + \left(\frac{\partial r}{\partial c}\right)x + \left(\frac{\partial s}{\partial c}\right),$$

so that $-(\partial r/\partial c)x - (\partial s/\partial c)$ is the remainder when $q(x)$ is divided by $(x^2 + bx + c)$. These remainders can be found by the method of § 9.55 (the quotients are also found, but are irrelevant), so the partial derivatives of r and s with respect to b and c can be determined.

If now changes Δb and Δc are made in b and c , the first-order changes in r and s are

$$\left(\frac{\partial r}{\partial b}\right)\Delta b + \left(\frac{\partial r}{\partial c}\right)\Delta c, \quad \left(\frac{\partial s}{\partial b}\right)\Delta b + \left(\frac{\partial s}{\partial c}\right)\Delta c,$$

and we want to choose Δb , Δc so as to reduce r and s to zero, that is, to make

$$r + \left(\frac{\partial r}{\partial b}\right)\Delta b + \left(\frac{\partial r}{\partial c}\right)\Delta c = 0, \quad s + \left(\frac{\partial s}{\partial b}\right)\Delta b + \left(\frac{\partial s}{\partial c}\right)\Delta c = 0.$$

These determine Δb , Δc and hence a better approximation

$$x^2 + (b + \Delta b)x + (c + \Delta c)$$

to the quadratic factor sought. The process can be repeated, and is second-order.

Example: To improve the approximation $x^2 + 1.970x + 1.965$ to a quadratic factor of $x^4 + 5x^3 + 12x^2 + 14x + 8$ (see example in previous section).

| Coefficient of | x^4 | x^3 | x^2 | x | 1 |
|----------------|--------------------------------|--------|---------|---------|---------|
| 1 | 1 | 5 | 12 | 14 | 8 |
| -1.970 | | -1.970 | -5.9691 | -8.0098 | |
| -1.965 | | | -1.965 | -5.9540 | -7.9895 |
| Quotient | 1 | 3.030 | 4.0659 | 0.0362 | 0.0105 |
| | $q(x) = x^2 + 3.030x + 4.0659$ | | | r | s |

| Coefficient of | x^3 | x^2 | x | 1 | x^2 | x | 1 |
|----------------|--------------|---|---|--------|------------|---|---|
| 1 | $x q(x) = 1$ | 3.030 | 4.066 | 0 | $q(x) = 1$ | 3.030 | 4.066 |
| -1.970 | | -1.970 | -2.088 | | | -1.970 | |
| -1.965 | | | -1.965 | -2.083 | | | -1.965 |
| | 1 | 1.060 | 0.013 | -2.083 | 1 | 1.060 | 2.101 |
| | | $-\left(\frac{\partial r}{\partial b}\right)$ | $-\left(\frac{\partial s}{\partial b}\right)$ | | | $-\left(\frac{\partial r}{\partial c}\right)$ | $-\left(\frac{\partial s}{\partial c}\right)$ |

Hence Δb , Δc are given by

$$0.013\Delta b + 1.060\Delta c = 0.0362,$$

$$-2.083\Delta b + 2.101\Delta c = 0.0105,$$

and solution of these equations gives $\Delta b = 0.0291$, $\Delta c = 0.0337$, whence

$$b = 1.970 + 0.0291 = 1.9991,$$

$$c = 1.965 + 0.0337 = 1.9987,$$

so that $x^2 + 1.9991x + 1.9987$ is a better approximation to a quadratic factor.

Actually the quadratic factors in this case are $x^2 + 2x + 2$ and $x^2 + 3x + 4$ exactly. One application of this method has improved the approximation to the factor $x^2 + 2x + 2$ by a factor of about 30.

9.6. Simultaneous non-linear equations

For simultaneous equations in two variables the same general procedure as for equations in one variable can be used, namely a graphical process for locating the roots approximately, followed by a numerical process for improving the approximation.

Let the equations be

$$f_1(x, y) = 0, \quad f_2(x, y) = 0. \quad (9.17)$$

If both of these can be solved formally for y as a function of x , or for x as a function of y , then it is easy to draw a graph of y against x for each equation, and the intersections of the two graphs give an approximation to the solutions. If one or both of the equations can be solved formally for y as a function of x or vice versa, then one of the variables can be eliminated and the equations reduced to an equation in one variable; for example if the second of equations (9.17) can be solved in the form $y = \phi_2(x)$, substitution of this in the first equation gives

$$F_1(x) \equiv f_1(x, \phi_2(x)) = 0.$$

There is no need to carry out the elimination explicitly in such a way as to exhibit $F_1(x)$ formally as a function of x ; all that is wanted is that $y = \phi_2(x)$ should be evaluated for a set of values of x , and that these should be substituted into the formula for $f_1(x, y)$ for the corresponding values of x . This process carries out the elimination numerically without it having to be expressed formally.

Example:

$$\sin x + 2 \sin y = 1,$$

$$2 \sin 3x + 3 \sin 3y = 0.3.$$

It is most convenient here to solve the first equation for $\sin y$, then from this to calculate $\sin 3y$ either from the formula

$$\sin 3y = \sin y(3 - 4 \sin^2 y)$$

or by use of inverse sine and sine tables, and then to evaluate

$$f_2(x) = 2 \sin 3x + 3 \sin 3y - 0.3$$

for these values of $\sin 3y$ and the corresponding values of $\sin 3x$. The work is conveniently arranged in tabular form.

| | $\sin y$ $= \frac{1}{2}(1 - \sin x)$ | $\sin 3y$ | $\sin 3x$ | $f_2(x)$ |
|----------------------------------|---|-----------|-----------|----------|
| 0 | 0.5 | +1 | 0 | +2.70 |
| $1(\frac{1}{8}\pi) = 30^\circ$ | 0.25 | 0.688 | 1 | 3.76 |
| $2(\frac{1}{8}\pi) = 60^\circ$ | 0.067 | 0.200 | 0 | +0.30 |
| $3(\frac{1}{8}\pi) = 90^\circ$ | 0 | 0 | -1 | -2.30 |
| $4(\frac{1}{8}\pi) = 120^\circ$ | 0.067 | 0.200 | 0 | +0.30 |
| $5(\frac{1}{8}\pi) = 150^\circ$ | 0.25 | 0.688 | 1 | 3.76 |
| $6(\frac{1}{8}\pi) = 180^\circ$ | 0.5 | 1 | 0 | +2.70 |
| $7(\frac{1}{8}\pi) = 210^\circ$ | 0.75 | +0.562 | -1 | -0.61 |
| $8(\frac{1}{8}\pi) = 240^\circ$ | 0.933 | -0.450 | 0 | -1.65 |
| $9(\frac{1}{8}\pi) = 270^\circ$ | 1 | -1 | 1 | -1.30 |
| $10(\frac{1}{8}\pi) = 300^\circ$ | 0.933 | -0.450 | 0 | -1.65 |
| $11(\frac{1}{8}\pi) = 330^\circ$ | 0.75 | +0.562 | -1 | -0.61 |
| $12(\frac{1}{8}\pi) = 360^\circ$ | 0.5 | 1 | 0 | +2.70 |

Two decimals are adequate to locate the roots approximately. A graph drawn from these values, or even inspection of the table without actually drawing a graph, shows that there are roots in the neighbourhood of $x/(\frac{1}{8}\pi) = 2.1, 3.9, 6.8,$ and 11.2 .

The approximate solutions so determined can be improved by tabulation at smaller intervals and inverse interpolation, or by an iterative process. If both the equations can be solved for one variable in terms of the other, say for y in terms of x :

$y = \phi_1(x)$ for the first equation,

$y = \phi_2(x)$ for the second equation,

then it may be more convenient to evaluate $\phi_1(x) - \phi_2(x)$ as a function of x and interpolate for the zero of this function.

Example: To find more exactly the root of

$\sin x + 2 \sin y = 1,$

$2 \sin 3x + 3 \sin 3y = 0.3$

in the neighbourhood of $x/(\frac{1}{8}\pi) = 2.1$

| $x/(\frac{1}{8}\pi)$ | x° | $\sin y$ $= \frac{1}{2}(1 - \sin x)$ | $y = \phi_1(x)$ | $\sin 3y$ $= 0.1 - \frac{2}{3} \sin 3x$ | $y = \phi_2(x)$ | $\phi_2(x) - \phi_1(x)$ | |
|----------------------|-----------|---|-----------------|--|-----------------|-------------------------|-----|
| 2.0 | 60° | .0670 | .0671 | .1 | .0334 | -.0337 | |
| | | | | | | | 240 |
| 2.05 | 61½° | .0606 | .0607 | .1523 | .0510 | -.0097 | -2 |
| | | | | | | | 238 |
| 2.1 | 63° | .0545 | .0545 | .2043 | .0686 | +.0141 | -5 |
| | | | | | | | 233 |
| 2.15 | 64½° | .0487 | .0487 | .2556 | .0862 | +.0374 | -2 |
| | | | | | | | 231 |
| 2.2 | 66° | .0432 | .0432 | .3060 | .1037 | +.0605 | |

and inverse interpolation then gives the required solution, approximately $x = 2.070(\frac{1}{8}\pi)$.

When neither of the equations can be solved formally for x or y , the same processes can be used, one or both of the functions $\phi_1(x)$, $\phi_2(x)$ being determined roughly graphically, or more accurately numerically, by solution of the equation $f_1(x, y) = 0$ or $f_2(x, y) = 0$ for y in terms of x . For example, if a set of graphs of $f_1(x, y)$ against y for a set of constant values of x is constructed, the intersections of these graphs with the y -axis give the function $y = \phi_1(x)$, which can then be used to substitute for y in the second equation.

Another process is to evaluate $f_1(x, y)$ and $f_2(x, y)$ for a set of points on a coarse grid in the (x, y) plane, and on a piece of squared paper to mark at each (x, y) point the values of $f_1(x, y)$ and $f_2(x, y)$ there. The loci $f_1(x, y) = 0$ and $f_2(x, y) = 0$ can then be sketched roughly, and the intersections of the curves thus sketched then indicate the regions of the plane in which a closer examination is necessary in order to determine the roots more accurately.

Example: To locate approximately the real solutions of

$$xy(2x^2 - y^2) + 16(x + y) = 48, \quad (9.18)$$

$$x^2 + y^2 = 16. \quad (9.19)$$

The second equation shows that x and y lie between ± 4 , so evaluate

$$xy(2x^2 - y^2) + 16(x + y)$$

on a square grid of mesh side unity in the (x, y) plane for $|x| \leq 4$, $|y| \leq 4$ (see Fig. 16). Although this grid is a coarse one, it enables the contour $f_1(x, y) = 0$, that is,

$$xy(2x^2 - y^2) + 16(x + y) = 48,$$

to be sketched roughly. In this particular case the contour $f_2(x, y) = 0$, that is, $x^2 + y^2 = 16$, could be drawn accurately; but in the figure it has been sketched freehand from the values of $x^2 + y^2$ at the mesh points, as would have to be done in general.

The intersections of the two contours show that there are four real solutions, approximately:

$$\begin{array}{cccc} x = -2.0 & x = 0.4 & x = 1.8 & x = 4.0 \\ y = 3.5 & y = 4.0 & y = 3.6 & y = -0.2 \end{array}$$

and probably two in the neighbourhood of $x = -3.5$, $y = -0.9$, though calculation of function values on a finer grid would be necessary in order to make certain whether the contours intersect in this region.

Notes: (i) The values recorded in the figure are not those of $f_1(x, y)$ and $f_2(x, y)$ themselves, but are those of the left-hand sides of equations (9.18), (9.19).

(ii) Since $f_1(x, y)$ in this case is a cubic in x for fixed y , and a cubic in y for fixed x , the table of its values can be checked both very easily and very thoroughly by differencing in both directions. Alternatively, these values could be built up from the differences in the y direction and checked by differencing in the x direction (or vice versa).

(iii) Since this method involves evaluating the function on a twofold array of points, it should in general be avoided when it is possible formally to use either

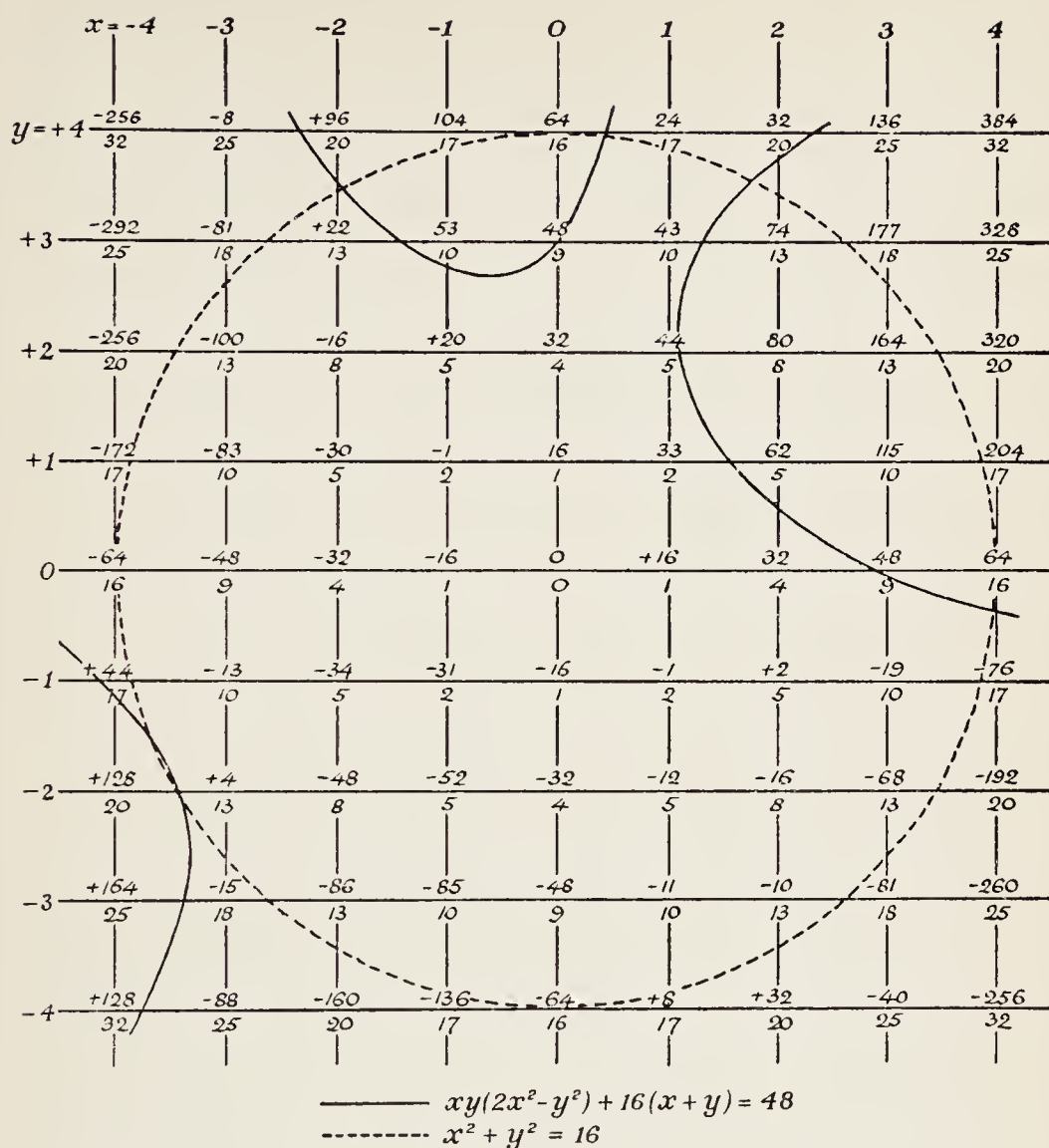


FIG. 16.

of the equations to solve for one or other variable in terms of the other, and so to reduce the problem to one in a single variable. It is used in this example to provide the possibility of comparing this method, and the results of using it and of improving the solutions, with others.

A convenient alternative method for these equations is to use the substitution $x = 4 \cos \theta$, $y = 4 \sin \theta$, which ensures that the equation (9.19) is satisfied, and then to treat equation (9.18) as an equation in θ . Another alternative is to use equation (9.19) to substitute for y^2 in (9.18), which then becomes

$$[(3x^2 - 16)x + 16]y = 48 - 16x, \quad (9.20)$$

and then treat equations (9.19) and (9.20) as two equations of the form $y = \phi(x)$.

(iv) The best method of improving the approximate solutions will depend on the equation and may be different for the different solutions. In the present

example an iterative process is most convenient for the solutions near $x = 0.4$, $y = 4.0$, and $x = 4.0$, $y = -0.2$. Consider the latter. We use equations (9.19) and (9.20) alternately, the first to determine x from an approximate value of y , and the second to determine y from an approximate value of x . Since $|y/x|$ is small for this solution, the value of x obtained from (9.19) is insensitive to the value of y taken, and is not much altered when an improved approximation to y , derived from (9.20) with this value of x , is used. The iterative process, though only first order, converges rapidly.

A similar treatment, using an equation obtained by substituting for x^2 instead of for y^2 in (9.18), is similarly effective for the solution near $x = 0.4$, $y = 4.0$.

The following is a general process for improving the approximate values of a solution, when neither equation can be solved formally for either variable in terms of the other. The functions $f_1(x, y)$, $f_2(x, y)$ are evaluated on a finer grid of points (x, y) in the neighbourhood of the solution. But instead of the function values being recorded in the (x, y) plane, the value of $f_2(x, y)$ is plotted against $f_1(x, y)$ for each pair of values (x, y) , and curves of constant x (x -contours) and of constant y (y -contours) are drawn in the (f_1, f_2) plane. The advantage of this method of representing the behaviour of the two functions of x and y is that each curve is drawn *through plotted points* instead of being interpolated 'by eye' among an array of function values.

Such a plot of f_2 against f_1 is made for such values of x and y that the point $f_1 = f_2 = 0$ is enclosed between two x -contours and two y -contours. Within a small enough region not containing more than one solution, the x -contours and y -contours will usually be nearly equally spaced and not very curved, and if this is the case, it is possible to estimate fairly closely what contours pass through the point $f_1 = f_2 = 0$. A calculation of (f_1, f_2) for this approximation to the solution then suggests for what further values of (x, y) the function should be evaluated in order to enclose the point $f_1 = f_2 = 0$ still more closely. The process is illustrated by the following example.

Example: To find more accurately the solution of the equations

$$xy(2x^2 - y^2) + 16(x + y) = 48, \quad x^2 + y^2 = 16,$$

in the neighbourhood of $x = 1.8$, $y = 3.6$.

From Fig. 16 it is estimated that the solution lies between $x = 1.7$ and 1.9 , and between $y = 3.5$ and 3.7 . The improvement of this solution can be carried out by the following procedure. A set of values of f_1 and f_2 is first evaluated for $y = 3.5$, 3.6 , and 3.7 , $x = 1.6$ to 2.0 , this range of x being taken in order to provide enough values to check by differences. The x -contours and y -contours drawn using these points are shown in Fig. 17; those for $y = 3.5$ and 3.6 already enclose the point $f_1 = f_2 = 0$; but some points for $y = 3.7$ have been calculated to check the spacing of the y -contours and to show the curvature of the x -contours. The (f_1, f_2) point

for $x = 1.85$, $y = 3.55$ is also shown. The solution estimated from these contours was $x = 1.84$, $y = 3.55$.

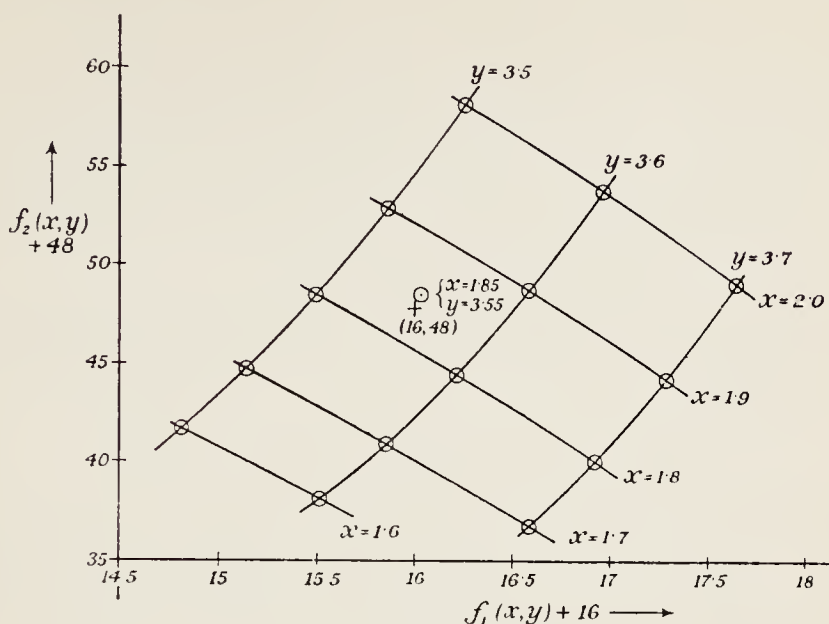


FIG. 17.

Values of (f_1, f_2) are now calculated for these four sets of values $x = 1.83, 1.84$ and $y = 3.55, 3.56$, and the results plotted on a larger scale. For this small range of x and y , the contours can be taken as straight and equally spaced to the accuracy of the plot. Inverse interpolation in x and y is required to determine the values of x and y to give $f_1 = f_2 = 0$, and this is most easily done by measurement. The values obtained can be checked by calculating f_1 and f_2 for them.

Notes: (i) As in Fig. 16, the functions plotted in Fig. 17 are the left-hand sides of the two equations, namely $f_1(x, y) + 48$ and $f_2(x, y) + 16$.

(ii) A convenient way of carrying out the final interpolation is as follows. Consider the interpolation between the two x -contours, say $x = x_0$ and x_1 , and let $x = x_0 + p(x_1 - x_0)$ be the interpolated value required. Lay a ruler on the (f_1, f_2) diagram so that its edge passes through the point $f_1 = f_2 = 0$; rotate it about this point and move it in the direction of its length until it cuts both x -contours at exact graduations on the scale, at a convenient interval (say 5, 10, or 20 units of the scale graduation); then the value of the fraction p of the x -interval between the contours can be read off directly.

(iii) This method of plotting is *not* satisfactory for the preliminary location of roots, since (x, y) is not in general a single-valued function of (f_1, f_2) , so that two or more x -contours and two or more y -contours may pass through each point in a region of the (f_1, f_2) plane. If this occurs, and it will occur if the equations have more than one solution, the (f_1, f_2) diagram becomes complicated and its interpretation needs considerable care.

9.7. Three or more variables

There is no satisfactory practical method, graphical or tabular, of displaying the behaviour of functions of three or more variables, and the approximate location of solutions of such equations is therefore difficult.

In some cases the solutions can be regarded as representing the asymptotic steady-state values of the solutions of a system of differential equations, and then they may be determined approximately by integrating this system of differential equations. If the equations arise from some scientific problem, this may suggest the appropriate differential equations to use. For example, in the chemical equilibrium of a system of a number of gaseous components, the relations between the concentrations of the components are given by a set of non-linear equations involving the equilibrium constants of the various reactions. If, for example, one of the reactions was $2\text{CO} + \text{O}_2 \rightleftharpoons 2\text{CO}_2$ and $\mathcal{C}(\text{X})$ stands for the concentration of the molecular species X, one equation would be

$$[\mathcal{C}(\text{CO})]^2 \mathcal{C}(\text{O}_2) = K_1 [\mathcal{C}(\text{CO}_2)]^2. \quad (9.21)$$

But the equilibrium is attained through a non-steady process in which the concentrations of the components change with time, that of oxygen, for example, being given by an equation

$$\frac{d}{dt} \mathcal{C}(\text{O}_2) = \beta_1 [K_1 \{\mathcal{C}(\text{CO}_2)\}^2 - \mathcal{C}(\text{O}_2) \{\mathcal{C}(\text{CO})\}^2], \quad (9.22)$$

and similarly for other components. We can try to make the calculations approach a steady state by following out such a time-varying process. However, since the purpose of the differential equation (9.22) is solely to provide a means for approaching a solution of the equation (9.21), there is no need to take experimental values of the reaction rate coefficients like β_1 in (9.22) even if these are known; an artificial set can be taken, convenient for the numerical work, and they need not even be taken to be constant.

X

FUNCTIONS OF TWO OR MORE VARIABLES

10.1. Functions of a complex variable and functions of two variables

THERE are two rather distinct contexts in which functions of two or more variables may arise in numerical work. One is concerned with complex numbers and functions of a complex variable. In numerical work it is usually best to treat a complex number as a pair of real numbers, either (x, y) in the Cartesian form $(z = x + iy)$ or (r, θ) in the polar form $(z = re^{i\theta})$ of the complex number as is most convenient for the calculation concerned. In this context a particularly important feature is the property of any analytic function $f(z) = g(z) + ih(z)$ of a complex variable z , that its real and imaginary parts both satisfy Laplace's equation in two dimensions. For this reason, the finite difference form of the two-dimensional Laplacian operator $(\partial^2/\partial x^2) + (\partial^2/\partial y^2)$ plays a particularly important part in such contexts.

The other is the general case of functions of two or more real variables other than those arising from formal expressions involving complex numbers. Here, too, the finite difference form of the Laplacian operator is important, particularly in two dimensions, and in three dimensions with some degree of spatial symmetry.

10.11. Numerical calculations with complex numbers

The details of numerical calculations with complex numbers will be carried out almost entirely with pairs of real numbers, since there is no standard calculating machine which deals directly with complex numbers. For addition and subtraction the Cartesian form $z = x + iy$ is clearly the more convenient. For multiplication and division the polar form $z = |z|e^{i\theta}$ seems preferable to the Cartesian form since although the Cartesian formulae

$$(x_1 + iy_1)(x_2 + iy_2) = x_1x_2 - y_1y_2 + i(x_1y_2 + x_2y_1),$$

$$(x_1 + iy_1)/(x_2 + iy_2) = [x_1x_2 + y_1y_2 + i(-x_1y_2 + x_2y_1)]/(x_2^2 + y_2^2)$$

are not difficult to evaluate, it is also not difficult to make a mistake of sign in this evaluation, particularly when x_1 , x_2 , y_1 , and y_2 are not all positive. Use of the polar form will probably involve some conversion from Cartesian to polar form; various good modern books of tables†

† For example, *Chambers's 6-Figure Tables*, vol. 2 (1949).

include tables for simplifying this conversion. Whether, and at what stages of a calculation, it is advisable to make a conversion from Cartesian to polar form or vice versa will depend so much on the calculation, and also perhaps on the individual worker and on whether Cartesian-polar conversion tables are available, that no general rule can be laid down.

For finding powers (other than squares and perhaps fourth powers) or roots of complex numbers, the polar form is usually the most convenient. But square roots can be found directly from the Cartesian form as follows. Let

$$(x+iy)^{\frac{1}{2}} = \xi + i\eta,$$

where ξ, η are real. On squaring and separating real and imaginary parts this gives

$$\xi^2 - \eta^2 = x, \quad 2\xi\eta = y. \quad (10.1)$$

Elimination of η gives a quadratic for ξ^2 , of which only the positive root is significant since ξ is real; this root is

$$\xi^2 = \frac{1}{2}\{x + |(x^2 + y^2)^{\frac{1}{2}}|\},$$

whence

$$\xi = [\frac{1}{2}\{x + |(x^2 + y^2)^{\frac{1}{2}}|\}]^{\frac{1}{2}}, \quad \eta = [\frac{1}{2}\{-x + |(x^2 + y^2)^{\frac{1}{2}}|\}]^{\frac{1}{2}}, \quad (10.2)$$

the signs of these square roots being taken so that $2\xi\eta = y$. If x is positive it may be best to use the first of formulae (10.2) to determine ξ , and then to find η from $\eta = y/2\xi$; and similarly if x is negative to use the second of formulae (10.2) to determine η , and then find ξ from $\xi = y/2\eta$. The result can be checked by squaring the value of $(\xi + i\eta)$ obtained.

10.2. Finite differences in two dimensions; square grid

Just as for functions of one variable x we often have to consider functions as specified by a table at discrete values of x , usually at equal intervals, so for a function of two or more independent variables we are often concerned with a function specified at discrete, equally spaced, values of all the independent variables. In particular, with two independent variables (x, y) it is very often most convenient to take these discrete values of x and y in such a way that they form a grid of square mesh in the (x, y) plane, such as

$$(x, y) = (x_0 + j\delta x, y_0 + k\delta y); \quad \delta x = \delta y = \delta s \quad (10.3)$$

with integral values of (j, k) . The values of a function f at such a point will be written $f_{j,k}$.

Such a function can be differenced in the x direction and in the y

direction; δ_x, δ_y will be used for central difference operators in the x and y directions, so that

$$\begin{aligned}\delta_x f_{j,k} &= f_{j+\frac{1}{2},k} - f_{j-\frac{1}{2},k}, & \delta_y f_{j,k} &= f_{j,k+\frac{1}{2}} - f_{j,k-\frac{1}{2}}; \\ \delta_x^2 f_{j,k} &= f_{j+1,k} - 2f_{j,k} + f_{j-1,k}, & \delta_y^2 f_{j,k} &= f_{j,k+1} - 2f_{j,k} + f_{j,k-1}.\end{aligned}$$

A particularly important relation is

$$(\delta_x^2 + \delta_y^2)f_{j,k} = f_{j+1,k} + f_{j,k+1} + f_{j-1,k} + f_{j,k-1} - 4f_{j,k}. \quad (10.4)$$

The operators $\delta x(\partial/\partial x)$ and $\delta y(\partial/\partial y)$ will be written U_x, U_y , the notation being an obvious extension of that of § 4.7. Then, as in § 4.7,

$$\delta_x = 2 \sinh \frac{1}{2} U_x, \quad \delta_y = 2 \sinh \frac{1}{2} U_y,$$

$$\text{and} \quad \delta_x^2 + \delta_y^2 = 2(\cosh U_x + \cosh U_y) - 4. \quad (10.5)$$

It is convenient to represent formulae such as (10.4), which represent linear combinations of values of f at a set of neighbouring points in the (x, y) plane, in a diagrammatic form in which the way in which the different function values enter is more immediately evident. Bickley† uses diagrams in which the set of coefficients in formula (10.4) would be represented by Fig. 18. A similar diagram, however, is used by Southwell in a different sense (see § 10.61, Fig. 19), and its use to represent the coefficients in formula (10.4) might be confusing. A more convenient form for printing is the following diagrammatic representation of formula (10.4):

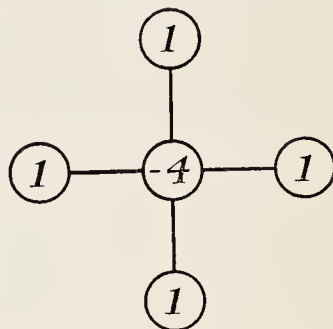


FIG. 18.

$$(\delta_x^2 + \delta_y^2)f_{j,k} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix} f_{j,k},$$

the set of coefficients being enclosed in a 'box' to distinguish it from a matrix. Another formula which will be needed, and which can be written in a similar form, is

$$\delta_x^2 \delta_y^2 f_{j,k} = \begin{bmatrix} 1 & -2 & 1 \end{bmatrix} \delta_y^2 f_{j,k} = \begin{bmatrix} 1 & -2 & 1 \\ -2 & 4 & -2 \\ 1 & -2 & 1 \end{bmatrix} f_{j,k}. \quad (10.6)$$

The quantity $(\delta_x^2 + \delta_y^2)f_{j,k}$ is four times the difference between the arithmetic mean of the values of f at the corners of a square centred on the point (j, k) and the value of f at the centre, the corners of the square

† W. G. Bickley, *Quart. J. Mech. and Applied Math.* **1** (1948), 35.

being the grid points which are the nearest neighbours of (j, k) (the side of the square is $\sqrt{2}(\delta s)$, not δs). A similar quantity involving next-nearest neighbours is

$$\begin{bmatrix} 1 & 0 & 1 \\ 0 & -4 & 0 \\ 1 & 0 & 1 \end{bmatrix} f_{j,k} = f_{j+1,k+1} + f_{j-1,k+1} + f_{j-1,k-1} + f_{j+1,k-1} - 4f_{j,k}.$$

In terms of the operators U_x, U_y this is

$$\begin{aligned} [e^{U_x+U_y} + e^{-U_x+U_y} + e^{-U_x-U_y} + e^{U_x-U_y} - 4]f_{j,k} \\ = [2 \cosh(U_x + U_y) + 2 \cosh(U_x - U_y) - 4]f_{j,k} \\ = 4[\cosh U_x \cosh U_y - 1]f_{j,k}. \end{aligned}$$

In terms of the differences of f it can be written:

$$\begin{bmatrix} 1 & 0 & 1 \\ 0 & -4 & 0 \\ 1 & 0 & 1 \end{bmatrix} f_{j,k} = \left\{ \begin{bmatrix} 1 & -2 & 1 \\ -2 & 4 & -2 \\ 1 & -2 & 1 \end{bmatrix} + \begin{bmatrix} 0 & 2 & 0 \\ 2 & -8 & 2 \\ 0 & 2 & 0 \end{bmatrix} \right\} f_{j,k} \\ = [\delta_x^2 \delta_y^2 + 2(\delta_x^2 + \delta_y^2)]f_{j,k},$$

$$\text{so that} \quad \delta_x^2 \delta_y^2 f_{j,k} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & -4 & 0 \\ 1 & 0 & 1 \end{bmatrix} f_{j,k} - 2(\delta_x^2 + \delta_y^2)f_{j,k}. \quad (10.7)$$

10.3. The operator $\partial^2/\partial x^2 + \partial^2/\partial y^2$

The particular importance of the Laplacian operator in two dimensions has already been noted in § 10.1. On a grid of square mesh of side δs we have

$$(\delta s)^2(\partial^2/\partial x^2 + \partial^2/\partial y^2) = U_x^2 + U_y^2,$$

and are therefore concerned with finite-difference approximations to $U_x^2 + U_y^2$. Using the approximations of § 4.71 we can express this in terms of the operators δ_x^2 and δ_y^2 as follows:

$$\begin{aligned} U_x^2 + U_y^2 &= \delta_x^2 [(\sinh^{-1} \tfrac{1}{2} \delta_x) / \tfrac{1}{2} \delta_x]^2 + \delta_y^2 [(\sinh^{-1} \tfrac{1}{2} \delta_y) / \tfrac{1}{2} \delta_y]^2 \\ &= \delta_x^2 - \tfrac{1}{12} \delta_x^4 + O(\delta x)^6 + \delta_y^2 - \tfrac{1}{12} \delta_y^4 + O(\delta y)^6 \\ &= \delta_x^2 + \delta_y^2 - \tfrac{1}{12}(\delta_x^4 + \delta_y^4) + O(\delta s)^6. \end{aligned} \quad (10.8)$$

Thus the simplest approximation to $U_x^2 + U_y^2$ is

$$U_x^2 + U_y^2 = \delta_x^2 + \delta_y^2 + O(\delta s)^4,$$

$$\text{which gives} \quad \left(\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} \right)_{j,k} = \frac{1}{(\delta s)^2} \begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix} f_{j,k} \quad (10.9)$$

with an error term of order $(\delta s)^2$. This approximation is widely used in numerical work. In particular it gives

$$\begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix} f_{j,k} = 0 \quad (10.10)$$

as a finite-difference form of $\partial^2 f / \partial x^2 + \partial^2 f / \partial y^2 = 0$.

10.31. Special relations when $\partial^2 f / \partial x^2 + \partial^2 f / \partial y^2 = 0$

In many contexts in which the operator $(\partial^2 / \partial x^2 + \partial^2 / \partial y^2)$ arises, its importance comes from the fact that one or more of the functions $f(x, y)$ concerned satisfy the relation

$$\partial^2 f / \partial x^2 + \partial^2 f / \partial y^2 = 0.$$

This is always the case when we are concerned with analytic functions of a complex variable, and is often the case in calculations not directly concerned with complex variables. If the operands are restricted to such functions, we have

$$U_x^2 + U_y^2 = 0,$$

and this can be used to obtain some special formulae for use in such contexts; but it must be remembered that they are restricted to such operands.

One of the most important can be derived as follows:

Since $U_x^2 + U_y^2 = 0$ it follows that

$$U_x^4 = U_y^4 = -U_x^2 U_y^2. \quad (10.11)$$

Hence

$$\delta_x^4 + \delta_y^4 = -2\delta_x^2 \delta_y^2 + O(\delta s)^6,$$

so that formula (10.8) can be written

$$0 = \delta_x^2 + \delta_y^2 + \frac{1}{6}\delta_x^2 \delta_y^2 + O(\delta s)^6,$$

and substitution from (10.7) gives

$$4(\delta_x^2 + \delta_y^2)f_{j,k} + \begin{bmatrix} 1 & 0 & 1 \\ 0 & -4 & 0 \\ 1 & 0 & 1 \end{bmatrix} f_{j,k} = O(\delta s)^6$$

that is,

$$\begin{bmatrix} 1 & 4 & 1 \\ 4 & -20 & 4 \\ 1 & 4 & 1 \end{bmatrix} f_{j,k} = 0 \quad (10.12)$$

with an error term of order $(\delta s)^6$. This is an improvement on the simplest finite-difference form (10.10) of the equation $\partial^2 f / \partial x^2 + \partial^2 f / \partial y^2 = 0$, for which the error is of order $(\delta s)^4$.

Another consequence of the relations (10.8) and (10.11) is that

$$\delta_x^4 = 6(\delta_x^2 + \delta_y^2) + O(\delta s)^6, \quad (10.13)$$

and this can, if convenient, be used in integration or interpolation formulae to substitute for fourth differences in the x direction in terms of the second differences. For example, one formula for integration in the x direction is

$$\int_{x_0 - \delta x}^{x_0 + \delta x} f dx = 2(\delta x)[f_0 + \frac{1}{6}\delta_x^2 f_0 - \frac{1}{180}\delta_x^4 f_0] + O(\delta x)^7 \quad (10.14)$$

(this is equivalent to Simpson's rule improved by the inclusion of the leading correcting term; see § 6.3). Expressed in diagrammatic form in terms of function values, this is

$$\int_{x_{j-1}}^{x_{j+1}} f(x, y_k) dx = \frac{1}{90}(\delta x)[-1 \quad 34 \quad 114 \quad 34 \quad -1]f_{j,k} + O(\delta x)^7. \quad (10.15)$$

Substitution for $\delta_x^4 f$ from (10.13) in (10.14) gives

$$\int_{x_0 - \delta x}^{x_0 + \delta x} f dx = 2(\delta x)[f_0 + \frac{2}{15}\delta_x^2 f_0 - \frac{1}{30}\delta_y^2 f_0] + O(\delta x)^7,$$

$$\text{that is, } \int_{x_{j-1}}^{x_{j+1}} f(x, y_k) dx = \frac{1}{15}(\delta x) \begin{bmatrix} 0 & -1 & 0 \\ 4 & 24 & 4 \\ 0 & -1 & 0 \end{bmatrix} f_{j,k} + O(\delta x)^7. \quad (10.16)$$

The coefficients are simpler in (10.16) than in (10.15) and the coefficient in the error term is smaller, as might be expected from the fact that the values of f involved in formula (10.16) lie nearer the range through which the integration is being carried than do the function values in (10.15).[†]

10.4. Finite differences in cylindrical coordinates

It is occasionally convenient to use finite differences at equal intervals in polar coordinates (r, θ) in a plane, or in cylindrical polar coordinates, rather than in Cartesian coordinates. Plane polar coordinates would be the natural ones to use, for example, in a calculation concerned with a solution of Laplace's equation in two dimensions with boundary conditions given on a circular boundary; and cylindrical polar coordinates would be the natural ones to use in a three-dimensional problem

[†] This formula was first derived by another method by G. Birkhoff and D. M. Young, see *Journ. of Math. and Phys.* **29** (1950), 217.

[‡] For a similar use of the relation $U_x^2 + U_y^2 = 0$ in the interpolation of functions of a complex variable, see P. M. and A. M. Woodward, *Phil. Mag.* (7) **37** (1946), 236; **39** (1948), 594.

with axial symmetry and boundary conditions on the surface of a circular cylinder. These cases can be considered together, the case of plane polar coordinates being given by putting $\partial/\partial z = 0$ in the equations for cylindrical polar coordinates.

One way of dealing with such calculations is to make the conformal transformation to $(\log r, \theta)$ and to work on a rectangular or square grid in the $(\log r, \theta)$ plane. But this is often not convenient when the point (or axis) $r = 0$ is in the domain to be covered by the integration, and it is then better to use the (r, θ) coordinates without modification.

Consider first the case of axial symmetry. Then the Laplacian operator in cylindrical polar coordinates is

$$\frac{\partial^2}{\partial r^2} + \frac{1}{r} \frac{\partial}{\partial r} + \frac{\partial^2}{\partial z^2}.$$

The finite-difference approximation to $\partial^2/\partial z^2$ is the same as in Cartesian coordinates; only the r -derivatives need special treatment. Let f_j stand for $f(j \delta r)$. Then

$$\left(\frac{\partial^2 f}{\partial r^2}\right)_j = (f_{j+1} - 2f_j + f_{j-1})/(\delta r)^2 + O(\delta r)^2.$$

An approximation to $(\partial f/\partial r)_j$, with an error term of the same order, is

$$\left(\frac{\partial f}{\partial r}\right)_j = (f_{j+1} - f_{j-1})/2(\delta r) + O(\delta r)^2,$$

so that for $j \neq 0$

$$\begin{aligned} \left(\frac{\partial^2 f}{\partial r^2} + \frac{1}{r} \frac{\partial f}{\partial r}\right)_j &= \left[\left(1 - \frac{1}{2j}\right)f_{j-1} - 2f_j + \left(1 + \frac{1}{2j}\right)f_{j+1}\right]/(\delta r)^2 + O(\delta r)^2 \\ &= [(2j-1)f_{j-1} - 4jf_j + (2j+1)f_{j+1}]/2j(\delta r)^2 + O(\delta r)^2. \end{aligned} \quad (10.17)$$

For axial symmetry, either there is a singularity at $r = 0$ or $\partial f/\partial r$ is zero there. If there is a singularity, further analytical investigation is required before numerical methods are applied. If $\partial f/\partial r = 0$ at $r = 0$, then

$$\left(\frac{\partial^2 f}{\partial r^2} + \frac{1}{r} \frac{\partial f}{\partial r}\right)_0 = 4(f_1 - f_0)/(\delta r)^2 + O(\delta r)^2, \quad (10.18)$$

a relation which can also be obtained from (10.9), since for axial symmetry each of the values of f with coefficient unity in (10.9) is f_1 .

If there is not axial symmetry, then there is an additional term $r^{-2}\partial^2/\partial\theta^2$ in the Laplacian operator, and if $f_{j,k}$ stands for $f(j \delta r, k \delta\theta)$, we have for $j \neq 0$,

$$\left(\frac{1}{r^2} \frac{\partial^2 f}{\partial \theta^2}\right)_{j,k} = \frac{1}{(j \delta r)^2} \left[\frac{f_{j,k+1} - 2f_{j,k} + f_{j,k-1}}{(\delta\theta)^2} \right] + O(\delta\theta)^2. \quad (10.19)$$

For the equation for $j = 0$, let \bar{f}_1 be the arithmetic mean of the values $f_{1,k}$ of f on the circle $r = \delta r$. Then

$$\left(\frac{\partial^2 f}{\partial r^2} + \frac{1}{r} \frac{\partial f}{\partial r} + \frac{1}{r^2} \frac{\partial^2 f}{\partial \theta^2}\right)_0 = 4(\bar{f}_1 - f_0)/(\delta r)^2 + O(\delta r)^2. \quad (10.20)$$

If f varies in the z direction, then to give $\nabla^2 f$, a finite-difference approximation to $\partial^2 f / \partial z^2$ has to be added to whichever of formulae (10.17) to (10.20) is the appropriate one to use for the variations in the (r, θ) plane.

10.5. Partial differential equations

Solutions of partial differential equations can sometimes be obtained by a separation of variables, by which the partial differential equation is reduced to a number of separate ordinary equations, one in each of the independent variables. Such a separation, if possible, is part of the preliminary analytical treatment of the problem before numerical methods come to be applied, and will not be considered here. The following sections are concerned with the numerical treatment of partial differential equations as such. It will mainly be concerned with partial differential equations in two independent variables, as the numerical solution of equations with three or more independent variables is usually a problem on too large a scale to handle without special equipment.

Many partial differential equations which arise in contexts in which numerical solutions are required are second order in at least one of the independent variables, and, moreover, are linear in the second-order derivatives. Simple examples are Poisson's equation in two dimensions

$$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = g(x, y), \quad (10.21)$$

where $g(x, y)$ is given; the equation of heat conduction or diffusion in one dimension

$$\frac{\partial f}{\partial t} = D \frac{\partial^2 f}{\partial x^2}, \quad (10.22)$$

in which the diffusivity D may depend on f (this would make the equation as a whole non-linear, but the second derivative enters linearly); and the wave equation

$$\frac{\partial^2 f}{\partial t^2} = \alpha^2 \frac{\partial^2 f}{\partial x^2}. \quad (10.23)$$

Just as the nature of the problem of numerical solution of ordinary differential equations depends on whether the conditions the solution has to satisfy are of the one-point or two-point type, so the nature of the

problem of the numerical solution of partial differential equations depends on whether the boundary conditions are given on a boundary completely enclosing the domain of the variables over which a solution is required, or whether this domain is unbounded in one or more directions. There is a classification of second-order equations in two variables as 'elliptic', 'parabolic', or 'hyperbolic' which is closely related to the different characters of boundary conditions usually associated with such equations, and the character of the problem of numerical integration is correspondingly different in the three cases.

The general second-order equation in two variables, linear in the second derivatives, is

$$H \frac{\partial^2 f}{\partial x^2} + 2K \frac{\partial^2 f}{\partial x \partial y} + L \frac{\partial^2 f}{\partial y^2} + M = 0, \quad (10.24)$$

where H , K , L , M may be functions of any one or more of the variables x , y , f , $\partial f/\partial x$, $\partial f/\partial y$. The classification depends on the sign of $K^2 - HL$; the reason for this will be explained in § 10.8. If this quantity is negative, the equation is termed 'elliptic'; if it is zero, the equation is termed 'parabolic'; and if it is positive, the equation is termed 'hyperbolic'. Poisson's equation (10.21) is a simple example of an 'elliptic' equation, the diffusion equation (10.22) is one of a 'parabolic' equation, and the wave equation (10.23) is one of a 'hyperbolic' equation. 'Elliptic' equations are usually associated with a domain completely bounded by closed curves (one of which may be the circle at infinity) on which boundary conditions are given. 'Parabolic' and 'hyperbolic' equations are usually associated with a domain which is open in the direction of one variable, which physically is often the time variable. For example we may require a solution of the heat conduction equation (10.22) from given *initial* conditions in time (f given as a function of x at $t = 0$) and with given *terminal* conditions in space (f given as a function of t at $x = a$, $x = b$) but with no condition to be satisfied at a later time $t = T$; the initial and terminal conditions are enough to define a solution, and such an independent condition at a later time could not generally be satisfied. It is not, however, *necessary* that the boundary conditions should be of this type; we might alternatively have no initial conditions, but given terminal conditions and a condition of periodicity in time, that is, a condition that f should be the same function of x at a given time T as at time $t = 0$.

If H , K , and L are not all constants, then the equation may be of different type in different parts of the domain in which the solution is

required. But for many of the simpler partial differential equations, such as (10.21), (10.22), and (10.23), including many practically important ones for which numerical work is likely to be needed, the equation remains of the same type throughout the whole domain, and only such cases will be considered here.

10.6. Elliptic equations

Poisson's equation (10.21) in two dimensions will be taken as a typical example of an elliptic equation for whose solution we require a numerical process. This process will cover as special cases Laplace's equation ($g(x, y) = 0$ in (10.21)) and the torsion equation ($g(x, y) = \text{const.}$). The first step is to replace the partial differential equation by a finite-difference relation on a convenient grid of discrete points. A Cartesian or polar grid will usually be most convenient, and for the present only a Cartesian grid of square mesh with mesh side h will be considered. It is clearly most convenient if the boundaries are of such a form that a grid can be chosen so that the boundaries lie along the sides or diagonals of the grid squares, and it will be supposed for the present that this is the case and that the grid is so chosen.

Using the simplest approximation (10.9) to $[(\partial^2/\partial x^2) + (\partial^2/\partial y^2)]$ we then have a set of equations

$$\begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix} f_{j,k} = h^2 g_{j,k}, \quad (10.25)$$

one for each mesh point. These are linear simultaneous algebraic equations, so that we have formally reduced the numerical problem to one of the kind already considered in Chapter VIII. The solution of the set of equations (10.25) is not, of course, the solution of the partial differential equation on account of the truncation error of the approximation (10.9). The approximation can be improved by taking a finer mesh or by using the better approximation (10.8) to $\partial^2 f/\partial x^2 + \partial^2 f/\partial y^2$. If the latter process is used, a convenient procedure is to write the finite-difference equation

$$\begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix} f_{j,k} = h^2 g_{j,k} + \frac{1}{12}(\delta_x^4 + \delta_y^4)f_{j,k} \quad (10.26)$$

and to solve this by an iterative process, using in the n th stage of the

iterative process values of $(\delta_x^4 + \delta_y^4)f$ obtained from the results of the $(n-1)$ th stage.†

10.61. Relaxation process

A 'relaxation' process (§ 8.5) is very convenient for carrying out the numerical solution of the set of equations (10.25) or (10.26), and is commonly used for this purpose. This common association of the relaxation procedure with the approximate equations (10.25) seems to have given the impression that the relaxation process itself is approximate, and the errors of the approximation (10.25) are sometimes referred to as 'errors of the relaxation process'. But this is a misunderstanding; the approximation is not in the relaxation process itself but in the equations (10.25) whose solution is evaluated by this process. Regarded as a solution of the partial differential equation, the solution of equations (10.25) is equally in error whether it is evaluated by the relaxation process or by any other (such as elimination or inversion of the matrix of the coefficients of equations (10.25)) and the errors have nothing to do with the relaxation process used to obtain a solution of these finite-difference equations.

The approximation to the solution of the partial differential equations can be improved by reducing the mesh size of the grid on which the finite differences are taken. In practice it is advisable to start with a very coarse grid so that the number of grid points is quite small, and then to break down the grid to one of smaller mesh size when an approximate solution on the coarse grid has been reached. Then the relaxation process on the finer grid starts from a set of values which is already a fair approximation to the solution.

It is convenient to carry out the relaxation process on a diagram representing the domain in which the solution is required, with the finite-difference grid drawn on it. The usual convention is to write the function values and their *changes* to the left of each grid point, and *values* of the residuals to the right. For the simplest finite-difference approximation (10.25) to Poisson's equation, the residual $R_{j,k}$ at the point (j, k) is

$$R_{j,k} = \begin{vmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{vmatrix} f_{j,k} - h^2 g_{j,k}. \quad (10.27)$$

If a relaxation Δf is made at one point, the residual at that point is changed by $-4\Delta f$, and that at each nearest neighbour, other than a

† See, for example, L. Fox, *Proc. Roy. Soc. A*, **190** (1947), 31.

boundary point, is changed by $+\Delta f$, so that the pattern of the *changes in the residuals* is as represented diagrammatically in Fig. 19. The entries here are the coefficients of a *single* Δf value.

Example: To find approximately the solution of Laplace's equation

$$\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} = 0$$

for the system shown in Fig. 20, with equipotentials $V = 0$ and $V = 80$ as indicated.

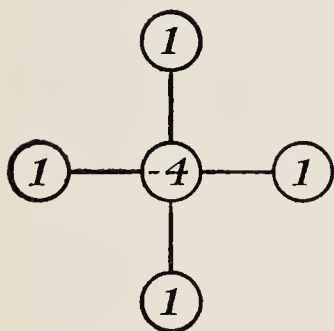


FIG. 19.

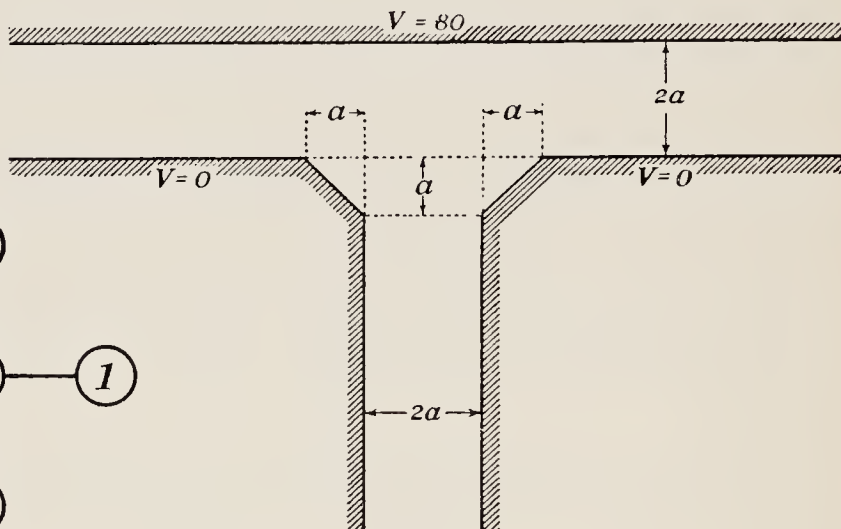


FIG. 20.

The first, coarse, grid can be taken as shown in Fig. 21. It might seem at first sight that this grid is too coarse for the results to be of any value. But we shall see that this rough approximation is in fact useful, and is obtained much more easily and quickly than results on a finer grid. By symmetry, only half of the diagram need be shown, but it must be remembered that each relaxation ΔV at a point one interval from the centre line is accompanied by an equal one at the image point, so that the contribution to the residual on the centre line is $2\Delta V$.

A set of values of V from which to start the relaxation process can be written in as if the equipotential $V = 0$ were the straight line AB . Then the residuals are zero except on AB . These values of V and the residuals are entered on Fig. 21.

We could start the relaxation process by making such a relaxation as to reduce the residual at C (for example) to zero; this would require a relaxation $\Delta V = +10$ at C (and at its image in the centre line). But clearly a positive relaxation ΔV is going to be required at D , which will make a positive contribution to the residual at C , and a further positive relaxation ΔV at C will be needed to remove it. We can anticipate this by deliberately taking a larger relaxation ΔV at C than is required to reduce the residual there to zero; this is called 'over-relaxing'. Experience is the only way of learning when and by how much to over-relax; the beginner will probably be inclined not to over-relax enough. As a rough rule it may be suggested that when there are several residuals of the same sign together, over-relaxation by a factor 2 will not be excessive.

values of V and the corresponding residuals. This check should normally be made before the accuracy of the numerical work is increased by taking an extra significant figure (compare the examples in §§ 8.5 and 8.51) and always before changing from a coarser to a finer grid.

Notes: (i) The individual numerical steps of the relaxation process are very simple and are carried out with small numbers, usually of one or two significant figures; they can therefore be carried out rapidly and easily.

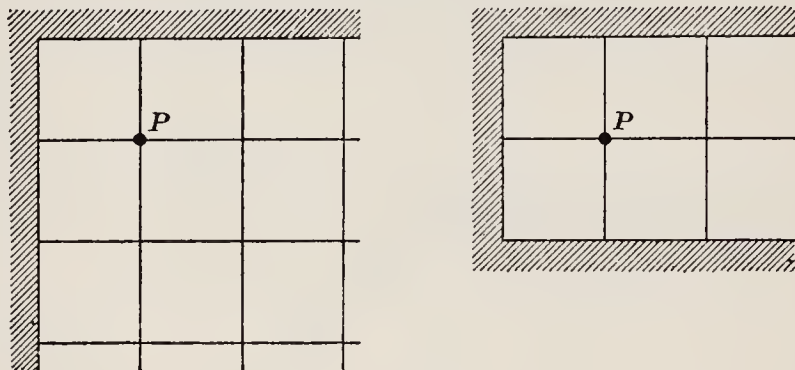


FIG. 22.

(ii) When the value of V at a point *not* at a distance δs from a boundary is changed, the sum of the residuals remains unchanged; all that is changed is the distribution of this total among the grid points. But if a relaxation ΔV is made at a grid point P adjacent to a boundary, the sum of the residuals is reduced by ΔV , or by $2\Delta V$ or $3\Delta V$ if two or three of the nearest neighbours of P are on the boundary (see Fig. 22).

(iii) A physical analogue of the relaxation process, as applied to the finite-difference form of Laplace's equation, can be given by considering Poisson's equation for the potential of a two-dimensional distribution of electrical charge, namely

$$\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} = -4\pi\rho.$$

The finite-difference approximation (10.9) to the left-hand side gives

$$R_{j,k} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix} V_{j,k} = h^2 \left[\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} \right] = -4\pi h^2 \rho_{j,k}, \quad (10.28)$$

and $h^2 \rho_{j,k}$ is (to this approximation) the charge on a square of side h centred on the point (j, k) . Thus, for any assigned set of values of $V_{j,k}$, the residuals $R_{j,k}$ are a measure of the charge distribution required to give the assigned potential distribution. The relaxation process can be regarded as a process of shifting this charge distribution about until it is ultimately all in the form of surface charge on conductors forming the given equipotential boundaries and none is left as space charge in the domain over which the integration is carried.

The constancy of the sum of the residuals when a relaxation is made at a grid point *not* adjacent to a boundary corresponds to the constancy of the total space charge in the domain when some charge is taken from one grid point and distributed among its four nearest neighbours. The change in the sum of the residuals when a relaxation is made at a point adjacent to a boundary corresponds to the transfer of some of the space charge to surface charge on the boundary.

This analogy suggests that the aim of the relaxation process should be not only to make the residuals small but to make them not all of the same sign, so that their sum, represented by the total residual space charge in this analogy, is small. It will not in general be possible to reduce all residuals to zero in the least significant digital position; a sprinkling of values ± 1 with occasional values ± 2 is the best that can be expected, and such a set of residuals, with mean value perhaps 0.1 or 0.2, probably indicates a better approximation to a solution than a set of residuals ± 1 over the whole field.

(iv) A set of residuals of magnitude not greater than 2 does not necessarily mean that the values of V are correct to a unit. It is advisable to reduce the residuals on the final grid to ± 2 in the next figure beyond the last figure in V required in the final results.

(v) No indications such as the letters (a), (b), (c),... in Fig. 21 are required in actual working; they are only given in this figure to help the reader to follow the details of the calculation. As soon as one value of a residual is replaced by another, the earlier one can be crossed out or erased as being of no further interest.

(vi) If the grid is drawn in ink and the working is done lightly in pencil, then old values of V and old residuals can be erased without losing the pattern of the grid. This erasing of old values need not be done at every relaxation, but only when the space for values of residuals gets filled up.

(vii) In this example the over-relaxation by a factor of 2 in the first relaxation has been a little too much, and a small relaxation of the opposite sign has had to be made later. But this step of over-relaxation has speeded the approach to a solution of the finite-difference equation. Only fourteen steps of relaxation have been needed to reduce the greatest value of $|R_{j,k}|$ from 40 to 2.

(viii) With the very coarse grid used here, there is no point in trying to improve the approximation to the solution of the finite-difference equations by taking an extra figure in the V -values. The next step is to reduce the truncation errors by taking a finer grid.

10.62. Reducing the mesh size

At some stage in the calculation it will usually be necessary, as in the above example, to change from a coarse to a finer grid. Let h_1 be the mesh size of the coarser grid. A convenient first step is to take the diagonals of the squares of the old grid as forming a new grid of mesh size $h_2 = h_1/\sqrt{2}$ (see Fig. 23). The new grid points are the centres of the squares of the old grid. For Poisson's equation we have on the new grid, with the finite-difference approximation adopted

$$\begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix} f_{j,k} = h_2^2 g_{j,k} = \frac{1}{2} h_1^2 g_{j,k},$$

$$\text{and hence} \quad f_{j,k} = \frac{1}{4} \left\{ \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} f_{j,k} - h_2^2 g_{j,k} \right\}. \quad (10.29)$$

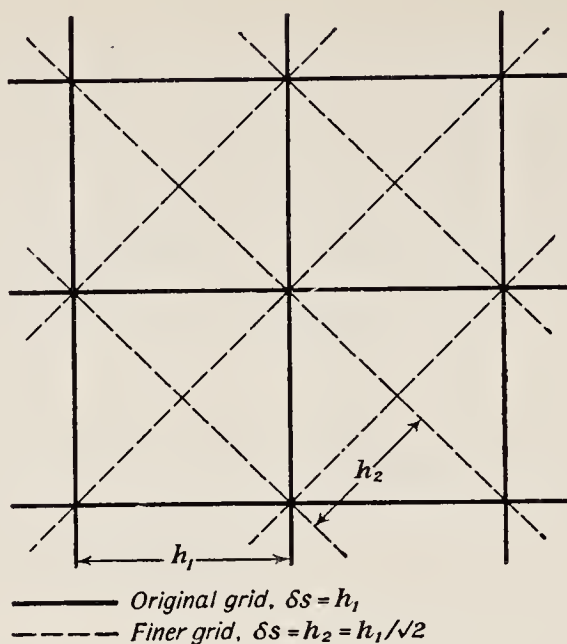


FIG. 23.

This gives a set of values of f at the centres of the squares of the old grid, which are the grid points of the new grid. In particular for Laplace's equation we have, in this approximation

$$f_{j,k} = \frac{1}{4} \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} f_{j,k}; \quad (10.30)$$

that is to say, the value of f at the centre of a square is the arithmetic mean of its values at the corners. A further relaxation may be carried out on this grid, or this process may be repeated immediately, giving a grid of mesh size $h_3 = \frac{1}{2}h_1$, whose grid points are the corners, centres, and mid-points of the sides of the original grid (see Fig. 24).

Example: The example of the previous section continued. Fig. 25 shows the process of breaking down the grid in two stages. The numbers in squares are the values of V obtained in the calculation shown in Fig. 21. The numbers in circles, at the centres of the squares of the original grid, are obtained by the application of formula (10.30) to the intermediate grid formed by the diagonals of the original grid. The numbers at the other grid points of Fig. 25 are then obtained by the application of formula (10.30) to the grid formed by the diagonals of the intermediate grid.

The residuals are shown on the right of the grid points, and only a single step of relaxation is then required to reduce the greatest $|R_{j,k}|$ to 2. At this stage another significant figure can be taken in V and the relaxation process continued.

Note: The advantage of starting with a very coarse grid will now be apparent. The number of grid points varies as $1/h^2$ and the number of relaxations at each grid point probably varies roughly as $1/h$, so that if the finer grid of Fig. 25 had been used from the beginning, something like eight times as much work would be required to reach the stage represented by the results in Fig. 25. In terms of the analogy explained in note (iii) of the previous section, despite the coarse grid of Fig. 21, relaxation on this grid has carried out the bulk of the transfer of charge from the inter-electrode space to the electrodes, and what has to be done on the finer grid is mainly a minor rearrangement of the residual charges.

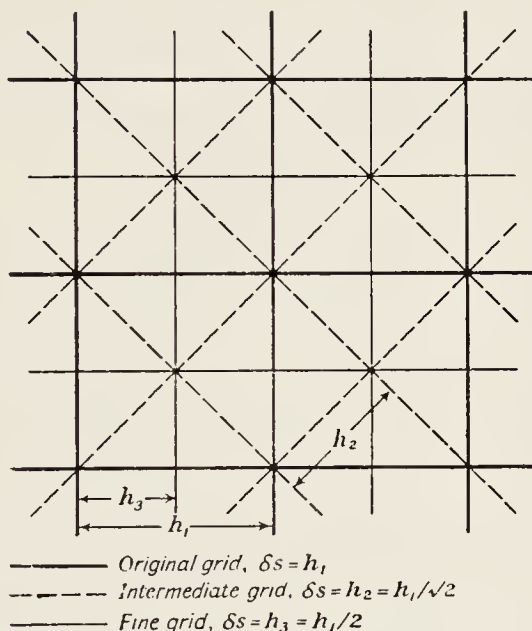


FIG. 24.

10.63. Further notes on the relaxation process

We have only been concerned here with the simplest case in which (i) the boundary of the domain of integration does not cut the side of any of the grid squares, (ii) the boundary condition is that V is given, and (iii) the equation to be solved is the simplest example of an elliptic equation. For extensions of the procedure to deal with boundaries which cut the sides of some of the grid squares, with boundary conditions involving the normal derivative of V , and with less simple equations, for further practical hints on carrying out the relaxation process in this context, and for examples, reference should be made to Southwell's *Relaxation Methods in Theoretical Physics* and papers referred to in the bibliography in that book.†

† See also E. Stiefel, *Zeit. f. angew. Math. und Phys.* **3** (1952), 1.

solution. Use of the finite-difference approximation (10.9) on the left-hand side gives the set of simultaneous equations

$$\begin{bmatrix} 0 & 1 & 0 \\ 1 & -(4-\lambda h^2) & 1 \\ 0 & 1 & 0 \end{bmatrix} f_{j,k} = 0,$$

and the determination of λ by a relaxation process follows the general lines of § 8.73.

10.64. Richardson–Liebmann process for Laplace’s equation

There is another process of successive approximation for solving the set of equations (10.10) which form the simplest finite-difference approximation to

$$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = 0.$$

In its simplest form, given by Liebmann,[†] this process consists of repeated use of formula (10.30), working systematically over the grid, replacing f at each grid point by the arithmetic mean of the value of f at its four nearest neighbours.

In another form, given by L. F. Richardson,[‡] each value of $f_{j,k}$ in a trial solution is increased by a multiple α of the residual $R_{j,k}$ at that point, and the result is taken as the next trial solution. Richardson proposed the use of a set of different values of α in the construction of successive trial solutions. Liebmann’s process is equivalent to a special case of Richardson’s in which α is kept fixed.

Compared with the more recently devised relaxation process, the Richardson–Liebmann process has three disadvantages. First, all the work is done with large numbers, the values of f themselves, whereas in the relaxation process the bulk of the work is done with relatively small and simple numbers, the *relaxations* of f and the residuals. Secondly, a lot of time and work is spent on calculation in regions where the residuals are small, whereas in the relaxation process attention is first directed to the region where the residuals are large and the rest of the domain is left untouched until the larger residuals have been removed. And, thirdly, it is not so easy to modify so as to take into account the higher differences in the replacement of derivatives by finite differences.

10.7. Parabolic equations

Most work on the numerical solution of parabolic partial differential equations has been concerned with the equation of heat conduction or

[†] H. Liebmann, *Sitzungsber. Bayer. Akad. München* (1918), 385.

[‡] L. F. Richardson, *Phil. Trans. Roy. Soc. A*, **210** (1910), 307.

diffusion.† Some practical methods are indicated in the following sections (§§ 10.71–10.73).

As a simple case of a parabolic equation we will consider the equation of heat conduction in one dimension

$$\frac{\partial f}{\partial t} = \frac{\partial^2 f}{\partial x^2}, \quad (10.31)$$

with given initial and terminal conditions.

One way of dealing with this equation is first to replace only one of the derivatives by a finite difference; this replaces the partial differential equation by a set of ordinary equations which can then be treated by one of the methods of Chapter VII. The form of this set of ordinary equations and the process for their solution differ considerably according as it is the first-order (time) derivative or the second-order (space) derivative which is replaced by a finite difference.

10.71. Replacement of the second-order (space) derivative by a finite difference

Let $f_j(t)$ be written for the value of f at $x = j \delta x$ and at time t . Then replacement of the second derivative by a finite difference gives

$$\frac{df_j(t)}{dt} = [f_{j+1}(t) - 2f_j(t) + f_{j-1}(t)]/(\delta x)^2 + O(\delta x)^2. \quad (10.32)$$

This is a set of simultaneous first-order equations for the different functions $f_j(t)$ and these can be solved numerically without difficulty. The initial value of each f_j is given by the initial conditions. The truncation errors are of order $(\delta x)^2$; they can be estimated, and the leading term in the corrections applied, by Richardson's h^2 -extrapolation process (see § 7.51). This method is not restricted to one space variable and the time variable, and it is practicable to use it for the numerical solution of the equation of heat conduction in two space variables and, moreover, for a substance of which the thermal properties vary with temperature.‡

10.72. Replacement of the first-order (time) derivative by a finite difference

For a time interval δt , the time derivative at any value of x can be replaced by a finite difference as follows:

$$\left(\frac{\partial f}{\partial t}\right)_{x,t+\frac{1}{2}\delta t} = [f(x, t+\delta t) - f(x, t)]/(\delta t) + O(\delta t)^2,$$

† For a survey of applications to the diffusion equation, see J. Crank, *The Mathematics of Diffusion* (Clarendon Press, 1955).

‡ See N. R. Eyres and others, *Phil. Trans. Roy. Soc.* **240** (1946), 1.

and, with an error term of the same order, $\partial^2 f / \partial x^2$ at time $t + \frac{1}{2}\delta t$ can be replaced by the arithmetic mean of its values at the beginning and end of the time interval:

$$\left(\frac{\partial^2 f}{\partial x^2}\right)_{x, t+\frac{1}{2}\delta t} = \frac{1}{2} \left[\frac{\partial^2}{\partial x^2} \{f(x, t+\delta t) + f(x, t)\} \right] + O(\delta t)^2.$$

If the right-hand sides of these are equated and the error terms neglected, we have

$$\frac{\partial^2}{\partial x^2} [f(x, t+\delta t) + f(x, t)] = (2/\delta t)[f(x, t+\delta t) + f(x, t)] - (4/\delta t)f(x, t). \quad (10.33)$$

Given f as a function of x at time t , this is an *ordinary* differential equation for f as a function of x at time $t + \delta t$. There is a set of equations (10.33), one for each time interval. But they can be integrated *successively*, and do not have to be treated simultaneously as equations (10.32) do; the calculation proceeds interval by interval in t , the results $f(x, t + \delta t)$ for the end of one interval being the given function $f(x, t)$ for the beginning of the next.

In the integration of equation (10.33) it is not necessary to know the values of $\partial^2 f / \partial x^2$ at the beginning of the interval; the best procedure is to carry out the numerical solution regarding equation (10.33) as an equation for the quantity

$$u = [f(x, t+\delta t) + f(x, t)] \quad (10.34)$$

and then to subtract the known $f(x, t)$ to give $f(x, t + \delta t)$. If we write $(2/\delta t) = k^2$, equation (10.33) becomes

$$\frac{d^2 u}{dx^2} - k^2 u = -2k^2 f(x, t). \quad (10.35)$$

If two separate integrations covering the same range in t are carried out, with different time intervals δt , the leading term in the truncation error can be eliminated by Richardson's h^2 -extrapolation process (see § 7.51), and in many cases this will also correct for the next term in the truncation error.†

In this method we carry out successive integrations of a single equation (10.35) instead of simultaneous integrations of a set of equations, and, moreover, equation (10.35) is a second-order equation with the first derivative absent, which as mentioned in § 7.2 is the most convenient form of all for numerical treatment. However, the solution of this equation has to satisfy two-point boundary conditions in x , and a

† See D. R. Hartree and J. R. Womersley, *Proc. Roy. Soc. A*, **161** (1937), 363.

step-by-step integration of this equation as it stands may be difficult because of the extreme sensitiveness of the solution to initial conditions and to rounding errors, which is the more marked the smaller the value taken for the time interval δt . For this reason a more practical way of evaluating a solution is the process given in § 7.63 involving factorization of the operator $(d^2/dx^2 - k^2)$ in equation (10.35), or the matrix factorization procedure of Thomas and Fox (§ 8.6).†

For parabolic equations which are less simple than the simple diffusion equation (10.31), and particularly for non-linear equations, the two-point character of the boundary conditions and the sensitiveness of the direct step-by-step solution makes the method less straightforward than it may appear at first sight, and other precautions may be necessary in using finite-difference approximations to derivatives.‡

10.73. Replacement of both derivatives by finite differences

In the notation of § 10.71,

$$\left(\frac{\partial f}{\partial t}\right)_{x,t} = [f_j(t+\delta t) - f_j(t-\delta t)]/2\delta t + O(\delta t)^2$$

and
$$\left(\frac{\partial^2 f}{\partial x^2}\right)_{x,t} = [f_{j+1}(t) - 2f_j(t) + f_{j-1}(t)]/(\delta x)^2 + O(\delta x)^2.$$

These approximations invite us to equate the right-hand sides and so obtain (neglecting the error terms)

$$f_j(t+\delta t) = f_j(t-\delta t) - \{2\delta t/(\delta x)^2\}[f_{j+1}(t) - 2f_j(t) + f_{j-1}(t)].$$

This looks a very attractive formula, since if the solution has been carried to any value of t , it gives directly each value of $f_j(t+\delta t)$ separately in terms of known quantities, and a process of using this formula to integrate through successive intervals δt looks simple and straightforward. Unfortunately, however, such a process is unstable, and effects of rounding errors build up rapidly and uncontrollably.§

However, there is another way of using similar approximations which leads to a stable numerical process which is practicable but not quite so simple.§ This is based on equating approximations to $\partial f/\partial t$ and $\partial^2 f/\partial x^2$ not at grid points in the (x, t) plane but at points half-way in t between grid points. If in equation (10.33), $\partial^2 f(x, t)/\partial x^2$ is replaced by

† For an extension of this procedure to the treatment of a non-linear parabolic equation, see D. F. C. Leigh, *Proc. Camb. Phil. Soc.* **51** (1955), 320.

‡ For examples and further discussion, see D. R. Hartree, *Rep. and Mem. A.R.C.* Nos. 2426, 2427 (1939, issued 1949).

§ See J. Crank and P. Nicolson, *Proc. Camb. Phil. Soc.* **43** (1947), 50.

$[f_{j+1}(t) - 2f_j(t) + f_{j-1}(t)]/(\delta x)^2$, and a similar replacement is made for $\partial^2 f(x, t + \delta t)/\partial x^2$, the result is

$$f_{j+1}(t + \delta t) - \{2 + (\delta x)^2/(\delta t)\}f_j(t + \delta t) + f_{j-1}(t + \delta t) = -[f_{j+1}(t) - \{2 - (\delta x)^2/\delta t\}f_j(t) + f_{j-1}(t)]. \quad (10.36)$$

This is a set of simultaneous algebraic equations for $f_j(t + \delta t)$ as a function of x_j with boundary conditions of the two-point type in x_j ; they can be solved by an application of the relaxation process or by some other process of successive approximation.

10.74. Note on methods for parabolic equations

In all three of the methods considered in §§ 10.71 to 10.73 the process of evaluating an approximate solution is carried out in the direction of t increasing, t being in the conduction equation (10.31) the time variable, and in general that independent variable which does not occur in any second derivatives. All three methods are practicable only if the domain of integration is open in the direction of this variable, so that the whole solution does not have to satisfy any conditions at some later time in the course of the process of solution. As already mentioned in § 10.5 this is the most common situation with parabolic equations.

10.8. Hyperbolic equations. Characteristics

For hyperbolic equations methods similar to those for parabolic equations can be used, and in addition there is another class of methods peculiar to hyperbolic equations. These depend on the properties of sets of curves called 'characteristics' of a hyperbolic equation. As in § 10.5, let the equation be

$$H \frac{\partial^2 f}{\partial x^2} + 2K \frac{\partial^2 f}{\partial x \partial y} + L \frac{\partial^2 f}{\partial y^2} + M = 0, \quad (10.37)$$

where H, K, L , and M may be functions of any one or more of $x, y, f, \partial f/\partial x, \partial f/\partial y$, and consider the integration of the equation along a curve C in the (x, y) plane. Let ds denote an element of arc of the curve C , and d/ds a rate of change along C (see Fig. 26). Then for an element of arc ds of C

$$d\left(\frac{\partial f}{\partial x}\right) = \left(\frac{\partial^2 f}{\partial x^2} \frac{dx}{ds} + \frac{\partial^2 f}{\partial x \partial y} \frac{dy}{ds}\right) ds, \quad d\left(\frac{\partial f}{\partial y}\right) = \left(\frac{\partial^2 f}{\partial x \partial y} \frac{dx}{ds} + \frac{\partial^2 f}{\partial y^2} \frac{dy}{ds}\right) ds,$$

and hence

$$H \frac{dy}{ds} d\left(\frac{\partial f}{\partial x}\right) + L \frac{dx}{ds} d\left(\frac{\partial f}{\partial y}\right) = \left[\left(H \frac{\partial^2 f}{\partial x^2} + L \frac{\partial^2 f}{\partial y^2} \right) \frac{dx}{ds} \frac{dy}{ds} + \frac{\partial^2 f}{\partial x \partial y} \left\{ H \left(\frac{dy}{ds} \right)^2 + L \left(\frac{dx}{ds} \right)^2 \right\} \right] ds.$$

On substitution from the differential equation (10.37) this becomes

$$H \frac{dy}{ds} d\left(\frac{\partial f}{\partial x}\right) + L \frac{dx}{ds} d\left(\frac{\partial f}{\partial y}\right) = \left[-M \frac{dx}{ds} \frac{dy}{ds} + \frac{\partial^2 f}{\partial x \partial y} \left\{ H \left(\frac{dy}{ds} \right)^2 - 2K \frac{dx}{ds} \frac{dy}{ds} + L \left(\frac{dx}{ds} \right)^2 \right\} \right] ds. \quad (10.38)$$

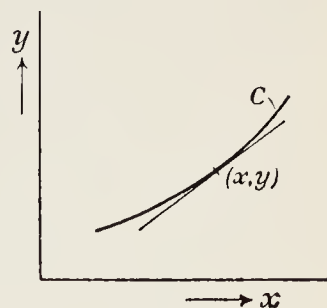


FIG. 26.

If now the curve C is chosen so that

$$H\left(\frac{dy}{ds}\right)^2 - 2K\frac{dx}{ds}\frac{dy}{ds} + L\left(\frac{dx}{ds}\right)^2 = 0, \quad (10.39)$$

then
$$H\frac{dy}{ds}\frac{d}{ds}\left(\frac{\partial f}{\partial x}\right) + L\frac{dx}{ds}\frac{d}{ds}\left(\frac{\partial f}{\partial y}\right) + M\frac{dx}{ds}\frac{dy}{ds} = 0. \quad (10.40)$$

Unless $dx/ds = 0$, these equations can be written

$$H\left(\frac{dy}{dx}\right)^2 - 2K\frac{dy}{dx} + L = 0, \quad (10.41)$$

$$H\frac{dy}{dx}\frac{d}{dx}\left(\frac{\partial f}{\partial x}\right) + L\frac{d}{dx}\left(\frac{\partial f}{\partial y}\right) + M\frac{dy}{dx} = 0. \quad (10.42)$$

A curve in the (x, y) plane such that equation (10.39) is satisfied at each point of it is called a *characteristic*. If $K^2 > HL$ (and only then), the roots of equation (10.39) at any point (x, y) are real and different, so that the characteristics are real; it is for this reason that the sign of $K^2 - HL$ is taken as the defining property to distinguish the classes of 'elliptic', 'parabolic', and 'hyperbolic' equations. Since for hyperbolic equations the roots of (10.39) are distinct it follows that through each point of the (x, y) plane there pass two characteristics. Thus there are two sets of characteristics covering the (x, y) plane, one member of each set passing through each point (x, y) . These two sets will be called 'set 1' and 'set 2'.

If H , K , and L do not depend on f , $\partial f/\partial x$, or $\partial f/\partial y$ (though they may depend on (x, y)), the characteristics are independent of the particular solution, and can be evaluated over the whole relevant domain of the (x, y) plane before the evaluation of a solution is started. But when one or more of H , K , and L depend on f , $\partial f/\partial x$, or $\partial f/\partial y$, the characteristics depend on the solution and the evaluation of the characteristics has to proceed simultaneously with that of the solution.

The essential feature of the characteristics, from the point of view of numerical integration of the equations, can be seen by comparing equation (10.40) with equation (10.38). In integration along any curve C , the integrand is a rate of change along C . Equation (10.40), as a first-order equation relating $\partial f/\partial x$ and $\partial f/\partial y$, involves only derivatives in this direction in the (x, y) plane; hence evaluation of the integrand for integration along a characteristic does not involve any differentiation in a direction across that in which the integration is being carried. On the other hand the presence of the term in $\partial^2 f/\partial x \partial y$ in equation (10.38) implies that in integration along a curve C other than a characteristic, the evaluation of the integrand for integration along C would involve differentiation in a cross direction. We have seen (§ 6.7) that numerical differentiation is a process which it is as well to avoid if possible; in the numerical solution of hyperbolic equations by integrating along curves in the (x, y) plane, it can be avoided if and only if these curves are taken to be the characteristics.

It is convenient to write μ_1, μ_2 for the roots dy/dx of (10.41), the value of μ_1 at any point referring to the characteristic of set 1 through that point, and the value of μ_2 to the characteristic of set 2. Then

$$\mu_1 + \mu_2 = 2K/H, \quad \mu_1 \mu_2 = L/H.$$

On a characteristic of set 1 we have

$$\frac{dy}{dx} = \mu_1, \quad (10.43)$$

$$\text{and from (10.42)} \quad \frac{d}{dx} \left(\frac{\partial f}{\partial x} \right) + \mu_2 \frac{d}{dx} \left(\frac{\partial f}{\partial y} \right) = -(M/H); \quad (10.44)$$

$$\text{on a characteristic of set 2,} \quad \frac{dy}{dx} = \mu_2, \quad (10.45)$$

$$\text{and} \quad \frac{d}{dx} \left(\frac{\partial f}{\partial x} \right) + \mu_1 \frac{d}{dx} \left(\frac{\partial f}{\partial y} \right) = -(M/H). \quad (10.46)$$

It sometimes happens that the derivatives $\partial f/\partial x$ and $\partial f/\partial y$, rather than f itself, are the quantities required in the solution, and further that H , K , L , M do not depend on f , though they may depend on $\partial f/\partial y$ and $\partial f/\partial x$; this is the case, for example, if f is the velocity potential of a steady isentropic irrotational flow of a compressible fluid, when $\partial f/\partial x$ and $\partial f/\partial y$, the components of the velocity, are the quantities really required. Then it is convenient to write u , v for $\partial f/\partial x$, $\partial f/\partial y$ respectively, and (10.44), (10.46) become

$$\frac{du}{dx} + \mu_2 \frac{dv}{dx} = -(M/H) \quad (10.47)$$

on a characteristic of set 1, and

$$\frac{du}{dx} + \mu_1 \frac{dv}{dx} = -(M/H) \quad (10.48)$$

on a characteristic of set 2.

10.81. Finite differences between characteristics

One way of adapting these equations for numerical work is, in effect, to use members of the two sets of characteristics as defining a finite-difference grid in the (x, y) plane and to work in terms of finite differences between neighbouring characteristics.† In Fig. 27 the two sets of curves represent the two sets of characteristics; the heavy portions represent the portions on which the solution has been carried out, and we want to determine the solution on the set of intersections of which A is typical.

On the characteristic AB of set 1, a finite-difference approximation to (10.43) is

$$\left. \begin{aligned} y_A - y_B &= \frac{1}{2}(\mu_{1A} + \mu_{1B})(x_A - x_B) \\ \text{and similarly on } AC \quad y_A - y_C &= \frac{1}{2}(\mu_{2A} + \mu_{2B})(x_A - x_C) \end{aligned} \right\} \quad (10.49)$$

Also on AB a finite-difference approximation to (10.47) is

$$(u_A - u_B) + \frac{1}{2}(\mu_{2A} + \mu_{2B})(v_A - v_B) = -\frac{1}{2}[(M/H)_A + (M/H)_B](x_A - x_B), \quad (10.50)$$

and similarly on AC ,

$$(u_A - u_C) + \frac{1}{2}(\mu_{1A} + \mu_{1C})(v_A - v_C) = -\frac{1}{2}[(M/H)_A + (M/H)_C](x_A - x_C). \quad (10.51)$$

The quantities u , v , x , y being known at B and C , this is a set of four equations for u , v , x , y at A .

If H , K , L , and so μ_1 and μ_2 , are independent of f , $\partial f/\partial x$, and $\partial f/\partial y$, then the first two equations give the position of A independently of the particular solution, and further, the coefficients on the left-hand sides of the second two equations are

† See, for example, L. H. Thomas, *Commun. on Pure and Applied Mathematics*, 7 (1954), 159.

known in advance of the solutions of these equations. The evaluation of a solution is then relatively simple. But if one or more of H , K , and L depend on f , $\partial f/\partial x$, or $\partial f/\partial y$, these four equations have to be solved as a set of simultaneous equations for u_A , v_A , x_A , y_A ; they are non-linear and can only be solved by trial and successive approximation. This makes the evaluation of a solution of the partial differential equation in such a case a long and often troublesome and tedious process.

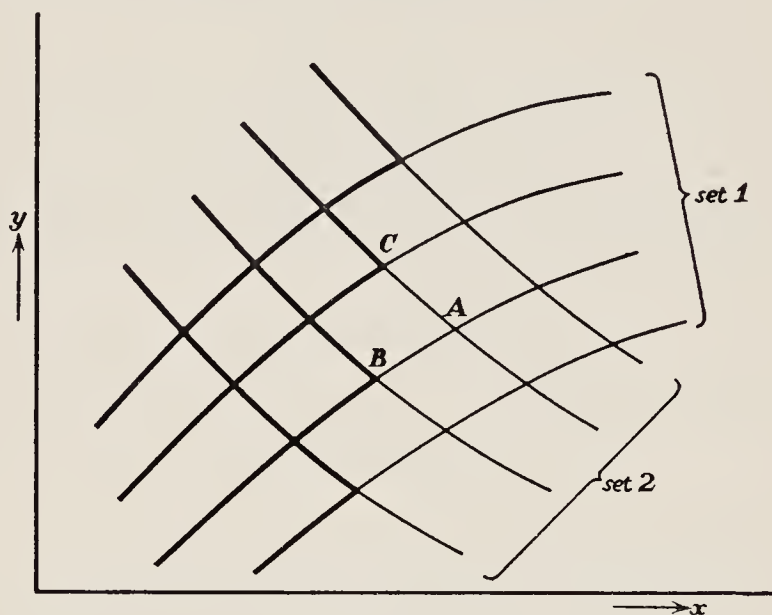


FIG. 27.

10.82. Use of given intervals in one independent variable

An alternative procedure, which has advantages in some contexts, is to use the relations (10.49), (10.50), and (10.51) which apply to characteristics, without using a grid of characteristics.

Suppose the solution is known on $x = x_0$, and consider a given interval δx to $x = x_0 + \delta x$. For any point A on the line $x = x_0 + \delta x$, let B and C be the points at which the characteristics through A cut the line $x = x_0$ (see Fig. 28). Then the relations (10.49), (10.50), and (10.51) of § 10.81 apply between the points A , B , and C , but what quantities are known and what quantities are unknown are now different. In § 10.81 the points B and C are given and the unknowns are x_A , y_A , u_A , and v_A ; now the point A is given and the unknowns are u_A , v_A , y_B , and y_C .

The fact that δx is known, and, moreover, is the same for all points A on the line $x = x_0 + \delta x$, can be used to simplify the solution of equations (10.49)–(10.51). The following is one procedure for finding this solution for a given point A . Equations (10.49) can be written

$$y_A - \frac{1}{2}\mu_{1A}\delta x = y_B + \frac{1}{2}\mu_{1B}\delta x, \quad (10.52)$$

$$y_A - \frac{1}{2}\mu_{2A}\delta x = y_C + \frac{1}{2}\mu_{2C}\delta x. \quad (10.53)$$

Now suppose $y + \frac{1}{2}\mu_1\delta x$, $y + \frac{1}{2}\mu_2\delta x$, μ_1 , μ_2 and (M/H) tabulated as functions of y for $x = x_0$. For given y_A , estimate u_A , v_A ; these give an approximation to μ_{1A} , and hence to $(y + \frac{1}{2}\mu_1\delta x)_B$ by (10.52). From the table of $y + \frac{1}{2}\mu_1\delta x$ as a function

of y , this gives y_B , μ_{2B} and $(M/H)_B$; and the values of x_A , y_A , u_A , and v_A give $(M/H)_A$. So (10.50) gives one relation between u_A and v_A . Similarly (10.53) gives an approximation to y_C , and (10.51) then gives another relation between u_A and v_A . These two relations can then be solved for u_A and v_A . If the estimates have been correctly made, the values of u_A and v_A so calculated will reproduce the estimates. One way of achieving this is to take three sets of estimates, namely $(u_A, v_A) = (a, b)$,

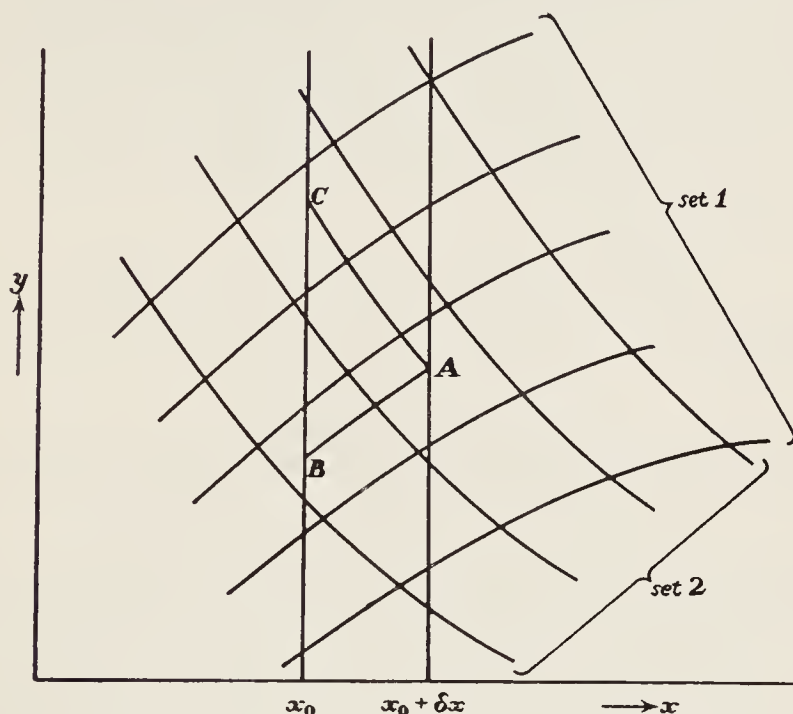


FIG. 28.

$(a + \delta a, b)$, and $(a, b + \delta b)$ and use linear interpolation to find the values of p_1 and p_2 for which estimates $(u_A, v_A) = (a + p_1 \delta a, b + p_2 \delta b)$ give results in agreement with the estimates; the calculation for this point A should be repeated with these estimates as a check. If the first estimate $(u_A, v_A) = (a, b)$ is not good, this process may have to be repeated.

This process may be simpler than the solution of equations (10.49)–(10.51) for x_A , y_A , u_A , and v_A , given B and C . Further it may give the results in a more convenient form, particularly if x is physically a time variable, for it gives results directly at exact values of x , which is the form in which they are likely to be required, whereas if results are obtained on a grid of characteristics, extensive interpolation is required to obtain a set of results for a set of values of x . Also the intervals at which results are obtained are completely under the control of the individual who is doing the work, instead of being determined by the shapes of the characteristics as the solution proceeds.

10.83. Two simultaneous first-order equations

For two simultaneous first-order equations, linear in the first derivatives, there may also be two sets of characteristic curves such that the evaluation of the

integrand along these curves involves no differentiation across them. Let the equations be

$$\frac{\partial f}{\partial x} + H_1 \frac{\partial f}{\partial y} + K_1 \frac{\partial g}{\partial y} = M_1, \quad (10.54)$$

$$\frac{\partial g}{\partial x} + H_2 \frac{\partial g}{\partial y} + K_2 \frac{\partial f}{\partial y} = M_2 \quad (10.55)$$

in which H_1, H_2, K_1, K_2, M_1 , and M_2 may be functions of one or more of the variables x, y, f , and g .

To find curves with the property required, multiply equation (10.55) by λ (which may be a function of x and y) and add equation (10.54). The result is

$$\left[\frac{\partial}{\partial x} + (H_1 + \lambda K_2) \frac{\partial}{\partial y} \right] f + \lambda \left[\frac{\partial}{\partial x} + (H_2 + \lambda^{-1} K_1) \frac{\partial}{\partial y} \right] g = M_1 + \lambda M_2. \quad (10.56)$$

In order that the operators

$$\left[\frac{\partial}{\partial x} + (H_1 + \lambda K_2) \frac{\partial}{\partial y} \right] \quad \text{and} \quad \left[\frac{\partial}{\partial x} + (H_2 + \lambda^{-1} K_1) \frac{\partial}{\partial y} \right]$$

should represent rates of change in the same direction in the (x, y) plane, it is necessary that

$$H_2 + \lambda^{-1} K_1 = H_1 + \lambda K_2, \quad (10.57)$$

that is,

$$\lambda^2 K_2 + \lambda(H_1 - H_2) - K_1 = 0.$$

If the roots λ_1, λ_2 of this equation are real and distinct, there are two such directions, given by

$$dy/dx = H_1 + \lambda_1 K_2, \quad dy/dx = H_1 + \lambda_2 K_2. \quad (10.58)$$

The pair of equations (10.54), (10.55) is then called 'hyperbolic' and the sets of curves on which dy/dx is given by the values (10.58) are called 'characteristics' of these equations (they are the characteristics, in the sense of § 10.8, of a second-order equation which can be derived from the two first-order equations).

For any function of x and y , say $h(x, y)$, $[\partial/\partial x + (H_1 + \lambda_1 K_1) \partial/\partial y]h$ is the rate of change of h with x along a curve C for which $dy/dx = H_1 + \lambda_1 K_2$. Hence on a curve C_1 , given by

$$(dy/dx)_1 = H_1 + \lambda_1 K_2, \quad (10.59)$$

equation (10.56) becomes

$$\left(\frac{df}{dx} \right)_1 + \lambda_1 \left(\frac{dg}{dx} \right)_1 = M_1 + \lambda_1 M_2; \quad (10.60)$$

and on a curve C_2 given by

$$(dy/dx)_2 = H_1 + \lambda_2 K_2 \quad (10.61)$$

equation (10.56) becomes

$$\left(\frac{df}{dx} \right)_2 + \lambda_2 \left(\frac{dg}{dx} \right)_2 = M_1 + \lambda_2 M_2. \quad (10.62)$$

These equations are very similar in form to equations (10.43) to (10.46) of § 10.8, and can be treated in a similar way. One important application is to the non-steady motion of a compressible fluid; in this case x is physically a time variable and y a space variable, and the roots of equation (10.57) are always real.

The relation between the characteristics of the two first-order equations (10.54) and (10.55), and those of a second-order equation can be shown as follows. Equation (10.57) expressed in terms of the values of $dy/dx = H_1 + \lambda K_2$ on the characteristic, is

$$\left(\frac{dy}{dx} \right)^2 - (H_1 + H_2) \frac{dy}{dx} + H_1 H_2 - K_1 K_2 = 0. \quad (10.63)$$

Also differentiation of equation (10.55) with respect to y , and of equation (10.54) with respect to x and to y , and elimination of second derivatives of g between the results, yields an equation in which the terms involving second derivatives are

$$\frac{\partial^2 f}{\partial x^2} + (H_1 + H_2) \frac{\partial^2 f}{\partial x \partial y} + (H_1 H_2 - K_1 K_2) \frac{\partial^2 f}{\partial y^2}$$

which are the second-order terms in an equation (10.37) in which

$$H = 1, \quad 2K = H_1 + H_2, \quad L = H_1 H_2 - K_1 K_2;$$

and for this differential equation, the equation (10.42) for the characteristics in the sense of § 10.8 is just equation (10.63).

XI

MISCELLANEOUS PROCESSES

11.1. Summation of series

IN practical applications of numerical analysis, as distinct from artificial examples constructed for the purpose, it is comparatively seldom that the original formulation of a problem is the summation of a series, though summation of a series is sometimes a useful method of dealing with a problem originally formulated in some other terms.

For example, the properties of the Airy function $\text{Ai}(x)$ which make it important in applications are these:

- (i) it is a solution of $y'' = xy$ which tends to zero as x tends to infinity; this defines it except for a constant multiplying factor;
- (ii) it is $\int_0^\infty \cos(xt + \frac{1}{3}t^3) dt$.

It can be evaluated from either of these properties without the use of its expansion as a power series in x . This power series expansion is a further property which happens to be useful in the evaluation of $\text{Ai}(x)$ for small values of x , but it is not the primary reason for the importance of this function, nor a property which need be used at all in its evaluation.

A series is useful in numerical work only if the sum of the first few terms is an adequate approximation to the sum of the series, or to the function represented by the series—just what a ‘few’ terms and an ‘adequate approximation’ mean will depend on the context. Suppose we have a numerical problem originally formulated in some other way than the summation of a series, and in trying to evaluate results by summing a series we find that the convergence of the first few terms is not rapid enough for them to be useful. Then this is a strong hint that evaluation of the series is not the best process for getting the results required, and the possibilities of other processes should be investigated.

But sometimes we may be concerned with the summation of slowly convergent series either in calculations originally formulated in such terms, or through the reduction of a more complicated situation to such a summation. In such cases we need processes for transforming slowly convergent series into more rapidly convergent ones. The simplest such transformation is one due to Euler for a series of terms of alternating signs.

11.11. Euler's transformation for a slowly convergent series of terms of alternate signs

This transformation can be derived by an application of finite-difference operators, and is one of the few cases in which the use of forward differences gives the most convenient form for results.

Let the *magnitudes* of the terms be u_0, u_1, u_2, \dots , in general u_n , so that the series which we wish to sum is

$$S = u_0 - u_1 + u_2 - u_3 + \dots = \sum_n (-1)^n u_n. \quad (11.1)$$

Let us take the successive differences of the terms u_n , regarded as a function of n . Also let Δ be the forward-difference operator with respect to n , defined by $\Delta u_n = u_{n+1} - u_n$, and $E = 1 + \Delta$. Then $u_n = E^n u_0$, and

$$\begin{aligned} S &= (1 - E + E^2 - E^3 + \dots) u_0 = \frac{1}{1 + E} u_0 = \frac{1}{2 + \Delta} u_0 = \frac{1}{2} (1 + \frac{1}{2} \Delta)^{-1} u_0 \\ &= \frac{1}{2} [u_0 - \frac{1}{2} \Delta u_0 + \frac{1}{4} \Delta^2 u_0 - \frac{1}{8} \Delta^3 u_0 + \dots]. \end{aligned} \quad (11.2)$$

The differences involved here are the *forward* differences from the first entry of the table of u_n , and are all available.

If the series (11.1) is slowly convergent, then the successive differences of the u_n 's usually decrease rapidly and the series (11.2) converges much more rapidly than the series (11.1). It will often be best not to carry the transformation back to the beginning of the series to be evaluated, but to calculate separately the sum of the first N terms ($N = 6$ or 8 , perhaps) and apply the Euler transformation to the remainder. A good check on the results can be obtained by carrying out this process with two different values of N .

Example: To calculate $S(x) = \sum_{m=0}^{\infty} (-1)^m / (x+m)^2$ for $x = 10$.

A table of the function $10^7/(x+m)^2$ for $x = 10$, $m = 0(1)10$ and its differences up to the sixth order is given on p. 266; the effects of rounding errors are becoming marked in the sixth differences. The value of $S(x)$ is the sum of *alternate* first differences of $-1/(x+m)^2$. If we take the first of these first differences (that is, the first two terms of the series) and apply the Euler transformation to the remainder of the series, we obtain

$$\begin{aligned} 10^7 S(10) &= 17355 + \frac{1}{2} [69444 + \frac{1}{2} (10272) + \frac{1}{4} (2120) + \frac{1}{8} (544) + \frac{1}{16} (162) + \\ &\quad + \frac{1}{32} (52) + \frac{1}{64} (14) + \dots] \\ &= 17355 + 37595 = 54950. \end{aligned}$$

The values of differences used are those underlined in the table.

| m | $10^7/(x+m)^2$ | | | | | | |
|-----|----------------|--------|-------------|-------|------------|------|-----------|
| 0 | 100000 | | | | | | |
| | | -17355 | | | | | |
| 1 | 82645 | | 4154 | | | | |
| | | -13201 | | -1225 | | | |
| 2 | <u>69444</u> | | 2929 | | 416 | | |
| | | -10272 | | -809 | | -151 | |
| 3 | 59172 | | <u>2120</u> | | 265 | | 48 |
| | | -8152 | | -544 | | -103 | |
| 4 | 51020 | | 1576 | | <u>162</u> | | 51 |
| | | -6576 | | -382 | | -52 | |
| 5 | 44444 | | 1194 | | 110 | | <u>14</u> |
| | | -5382 | | -272 | | -38 | |
| 6 | 39062 | | 922 | | 72 | | 19 |
| | | -4460 | | -200 | | -19 | |
| 7 | 34602 | | 722 | | 53 | | 0 |
| | | -3738 | | -147 | | -19 | |
| 8 | 30864 | | 575 | | 34 | | |
| | | -3163 | | -113 | | | |
| 9 | 27701 | | 462 | | | | |
| | | -2701 | | | | | |
| 10 | 25000 | | | | | | |

If we take the first four terms of the series and apply the Euler transformation to the remainder, we obtain

$$\begin{aligned}
 10^7 S(10) &= 17355 + 10272 + \\
 &\quad + \frac{1}{2}[51020 + \frac{1}{2}(6576) + \frac{1}{4}(1194) + \frac{1}{8}(272) + \frac{1}{16}(72) + \frac{1}{32}(19) + \dots] \\
 &= 27627 + 27323 = 54950.
 \end{aligned}$$

This agrees with the value already calculated, and we obtain the result $S(10) = 0.005495$ to six decimals.

If the ratios of successive terms u_{n+1}/u_n are nearly constant, a modified form of the Euler transformation can be used effectively. Let

$$v_n = \beta^n u_n,$$

β being a number chosen so that the variation of v_n with n is small. Then

$$\begin{aligned}
 S &= v_0 - \beta^{-1}v_1 + \beta^{-2}v_2 - \beta^{-3}v_3 + \dots \\
 &= [1 - (E/\beta) + (E/\beta)^2 - (E/\beta)^3 + \dots]v_0 \\
 &= \frac{1}{1 + (E/\beta)}v_0 = \frac{\beta}{(\beta+1) + \Delta}v_0 = \frac{\beta}{\beta+1} \left(1 + \frac{\Delta}{\beta+1}\right)^{-1}v_0 \\
 &= \frac{\beta}{\beta+1} \left[v_0 - \frac{1}{\beta+1}\Delta v_0 + \frac{1}{(\beta+1)^2}\Delta^2 v_0 - \frac{1}{(\beta+1)^3}\Delta^3 v_0 + \dots \right]. \quad (11.3)
 \end{aligned}$$

11.12. Use of the Euler-Maclaurin integration formula in the summation of series

When $f(x)$ is a function such that $\int f(x) dx$ can be integrated formally, the Euler-Maclaurin formula (6.22) can often be used effectively for

evaluating sums of the type $\sum_m f(m)$ over a set of integral values of m . From formula (6.22) with $x_0 = 0$ and interval $(\delta x) = 1$, we have

$$\begin{aligned} f_0 + f_1 + \dots + f_n \\ = \int_0^{x_n} f(x) dx + \frac{1}{2}(f_0 + f_n) + \frac{1}{12}(f'_n - f'_0) - \frac{1}{720}(f'''_n - f'''_0) + \frac{1}{30240}(f^{(v)}_n - f^{(v)}_0) - \dots \end{aligned} \quad (11.4)$$

and in particular, if $f(x)$ and all its derivatives tend to 0 as x tends to ∞ ,

$$\sum_{m=0}^{\infty} f_m = \int_0^{\infty} f(x) dx + \frac{1}{2}f_0 - \frac{1}{12}f'_0 + \frac{1}{720}f'''_0 - \frac{1}{30240}f^{(v)}_0 + \dots \quad (11.5)$$

As in the previous section, the result of using this formula can be checked by applying it to the series formed by omitting the first few terms from the series to be summed.

Example: To evaluate $\sum_{m=0}^{\infty} 1/(8+m)^2$.

For $f(x) = 1/(a+x)^2$, we have $\int_0^{\infty} f(x) dx = 1/a$,

and $f'(0) = -2/a^3$, $f'''(0) = -24/a^5$, $f^{(v)}(0) = -720/a^7$,

so evaluation of formula (11.5) for $a = 8$ gives

$$\begin{aligned} \sum_{m=0}^{\infty} 1/(8+m)^2 &= \frac{1}{8} + \frac{1}{2} \frac{1}{64} + \frac{1}{12} \frac{2}{8^3} - \frac{1}{720} \frac{24}{8^5} + \dots \\ &= .125 + .0078125 + .0003255 - .0000010 + \dots \\ &= .133137 \text{ to six decimals.} \end{aligned}$$

Also $\sum_{m=0}^{\infty} 1/(8+m)^2 = \frac{1}{64} + \frac{1}{81} + \sum_{m=0}^{\infty} 1/(10+m)^2$

and evaluation of formula (11.5) for $a = 10$ gives

$$\begin{aligned} \sum_{m=0}^{\infty} 1/(10+m)^2 &= \frac{1}{10} + \frac{1}{2} \frac{1}{100} + \frac{1}{12} \frac{2}{10^3} - \frac{1}{720} \frac{24}{10^5} \\ &= .1 + .005 + .0001667 - .0000003 \\ &= .1051664 \end{aligned}$$

so $\sum_{m=0}^{\infty} 1/(8+m)^2 = .015625 + .0123457 + .1051664$
 $= .133137$ to six decimals

verifying the value obtained by evaluating formula (11.5) with $a = 8$.

Slowly convergent series of positive terms which cannot be handled by this application of the Euler-Maclaurin formula are often difficult to deal with numerically. If the terms are given by an algebraical formula,

then it may be possible to find an analytical transformation which converts the series into a more rapidly convergent one, but this procedure is not usually available unless each term is of a relatively simple form.†

11.2. Harmonic analysis

Harmonic analysis is concerned with the representation of a function $f(x)$, over a finite range of x which will be taken as 2π , as a series of circular functions of x :

$$f(x) = \frac{1}{2}A_0 + A_1 \cos x + A_2 \cos 2x + \dots + B_1 \sin x + B_2 \sin 2x + \dots \quad (11.6)$$

The most important applications are to cases in which $f(x)$ is periodic in x with period 2π , or in which $f(x)$, although not periodic, or not defined outside a range $x_0 \leq x \leq (x_0 + 2\pi)$, satisfies the conditions

$$f^{(k)}(x_0 + 2\pi) = f^{(k)}(x_0). \quad (11.7)$$

If $f(x)$ does not satisfy such conditions, or if it has discontinuities in magnitude or in a differential coefficient of low order, then harmonic analysis is usually of formal rather than numerical interest, since in numerical work only a finite number of coefficients in the series (11.6) can be evaluated, and the sum of any finite number of terms gives a function which satisfies the conditions (11.7) and has no discontinuity in any derivative. If f itself has a discontinuity or does not satisfy

$$f(x_0 + 2\pi) = f(x_0),$$

then the behaviour of the sum of a finite number of terms in the neighbourhood of the discontinuity (or of x_0 and $x_0 + 2\pi$) differs considerably from the behaviour of $f(x)$; as n increases, the behaviour of f remains of the character shown in Fig. 29, the scale of x , but *not* the scale of the oscillations in f , becoming smaller as n increases. This is known as the ‘Gibbs phenomenon’ and illustrates the need for caution in regarding a finite number of terms of the series (11.6) as an adequate representation of the function unless it is free from discontinuities and the conditions (11.7) are satisfied.

The most usual applications of harmonic analysis are in connexion with the analysis of results of experiment or observation. Occasionally, however, it may be required in purely analytical or numerical contexts.

† For other methods of treatment of such series, see J. C. P. Miller and W. G. Bickley, *Phil. Mag.* (7) **22** (1936), 754; T. M. Cherry, *Proc. Camb. Phil. Soc.* **46** (1950) 436; G. G. Macfarlane, *Phil. Mag.* (7) **40** (1949), 188. See also O. Szász, *Journ. Math. and Phys.* **28** (1949), 272.

For example, in the solution of Laplace's equation in two dimensions,

$$\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} = 0,$$

in the interior of the unit circle, it may be convenient to use the result that V can be written

$$V = V_0 + r(A_1 \cos \theta + B_1 \sin \theta) + r^2(A_2 \cos 2\theta + B_2 \sin 2\theta) + \dots$$

Harmonic analysis of V as a function of θ on the unit circle gives V_0 and the coefficients A_n and B_n directly, and hence the solution V , without requiring numerical integration over the whole interior of the unit circle.

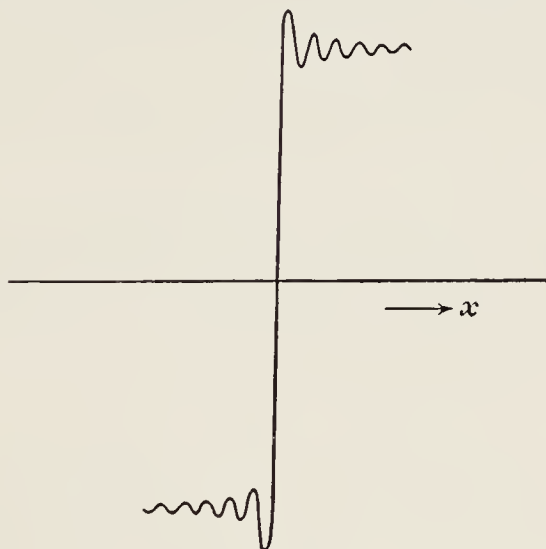


FIG. 29.

The coefficients in the series (11.6) are given by

$$\pi A_n = \int_0^{2\pi} f(x) \cos nx \, dx, \quad \pi B_n = \int_0^{2\pi} f(x) \sin nx \, dx. \quad (11.8)$$

If $f(x)$ satisfies the conditions (11.7) it follows that, in each of these integrals, each derivative of the integrand has the same value at the upper limit as at the lower limit. Hence in the Euler–Maclaurin formula (6.22) for each of the integrals the correcting terms from the two ends of the range cancel identically, and if the integrals can be evaluated from a set of values of $f(x)$ at equal intervals on x , the appropriate integration formula is the trapezium rule without corrections.† Hence if the range 2π in x is divided into K equal intervals, we have

$$\frac{1}{2} K A_n = \sum_{k=0}^{K-1} f(x_k) \cos nx_k, \quad \frac{1}{2} K B_n = \sum_{k=0}^{K-1} f(x_k) \sin nx_k, \quad (11.9)$$

where $x_k = 2\pi k/K$.

† See, however, the discussion of integrals of this kind in § 6.54.

These expressions for the coefficients are not significant for $n > \frac{1}{2}K$. This can be seen as follows. Let n_0 be a value of n less than $\frac{1}{2}K$, and m any positive integer. Then at the points $x_k = 2\pi k/K$ we have

$$\begin{aligned}\cos(mK \pm n_0)x_k &= \cos(2\pi mk \pm n_0 x_k) = \cos n_0 x_k, \\ \sin(mK \pm n_0)x_k &= \sin(2\pi mk \pm n_0 x_k) = \pm \sin n_0 x_k.\end{aligned}$$

Hence at these points the contributions from the terms with $n = mK \pm n_0$ have exactly the same variation with k as contributions from the term with $n = n_0$, and no analysis using only the values of f at these points can distinguish the contributions from the values $n = mK \pm n_0$ with different values of m . The first of the sums (11.9) gives the same values for each A_n ($n = mK \pm n_0$); but the values of the A_n 's are independent, so these values given by (11.9) cannot all be significant.

The point is that for values of n greater than $\frac{1}{2}K$ the values of x_k are not closely enough spaced, relative to the period of $\sin nx$, for the formulae (11.9) for the integrals to be valid. For $n = \frac{1}{2}K$ the values of the integrand $f(x)\cos nx$ at successive values of x_k are

$$+f(x_0), \quad -f(x_1), \quad +f(x_2), \quad -f(x_3), \quad \dots$$

and for most functions f these values are too irregular to give any confidence that they represent the behaviour of the integrand well enough to justify any numerical work on it at all. Their differences diverge and the situation is similar to that considered in § 6.54, where also we were concerned with an integral for which the correction to the trapezium rule vanished at both ends of the range of integration, but an incorrect value was obtained if too great an interval of integration was taken. To define an oscillating function adequately it is advisable to have at least six points per period, and this suggests that the series (11.9) should not be regarded as adequate approximations to the integrals (11.8) for $n > \frac{1}{6}K$.

Another aspect of these results can be illustrated by considering an alternative way in which the coefficients A_n , B_n might be determined numerically, not as values of the integrals (11.8) but by fitting the series (11.6) to $f(x)$ at the discrete set of values x_k of x . If this is done, then we have

$$\left. \begin{aligned}\sum_{k=0}^{K-1} f(x_k) \cos nx_k &= \frac{1}{2}K[A_n + A_{K-n} + A_{K+n} + A_{2K-n} + \dots] \\ \sum_{k=0}^{K-1} f(x_k) \sin nx_k &= \frac{1}{2}K[B_n - B_{K-n} + B_{K+n} - B_{2K-n} + \dots]\end{aligned}\right\}. \quad (11.10)$$

It is clear that with the restriction to K values of x , it is not possible to determine more than K relations between the coefficients, namely the first of relations (11.9) for $n = 0$ to $\frac{1}{2}K$ and the second for $n = 1$ to $(\frac{1}{2}K - 1)$; the terms $B_n \sin nx_k$ for $n = (m + \frac{1}{2})K$ make no contribution to the sum (11.6) at any of the points $x = x_k$. The smoothest function with the assigned values of $f(x_k)$ will be that for which $A_n = B_n = 0$ for $n > \frac{1}{2}K$, and then A_n, B_n for $n \leq \frac{1}{2}K$ are given by (11.9).

If $f(x)$ is a continuous function, then a good test of the significance of the values of A_n, B_n calculated from (11.9) is given by making two analyses with values of K which are relatively prime or have only a small common factor, such as $K = 30$ and 32 , or 48 and 50 ; this process also provides a good overall check on the results.

11.3. Recurrence relations for a sequence of functions

The Bessel functions of integral order $J_n(x)$ form an example of a set of functions of one variable (x) and one parameter (n) which have a number of properties in common, such as the form of the differential equation satisfied by them and their asymptotic behaviour. They are connected by relations between the functions of different orders n , such as

$$J_{n+1}(x) - J_{n-1}(x) = -2J'_n(x), \quad (11.11)$$

$$J_{n-1}(x) + J_{n+1}(x) = (2n/x)J_n(x). \quad (11.12)$$

Such relations are called recurrence relations. Other examples of such sets of function are the Legendre functions $P_n(x)$, the confluent hypergeometric functions $W_{k,m}(x)$ of Whittaker,[†] and the Weber functions $D_n(x)$.[‡]

It is often convenient to use such recurrence relations to evaluate functions, for some value of the parameter for which there may be no tables available, from tabulated values of the functions for other parameter values. Such a process must be used with care or it may lead to quite spurious results. This can be seen by considering, as an example, the evaluation of $J_n(x)$ for a given value of x and for large values of n from $J_0(x)$ and $J_1(x)$ by repeated use of the relation (11.12).

For $J_n(x)$ we require that solution of (11.11) which tends to zero as n tends to infinity. But if we evaluate $J_n(x), J_{n+1}(x), J_{n+2}(x), \dots$ in succession by using (11.12) in the form

$$J_{n+1}(x) = (2n/x)J_n(x) - J_{n-1}(x), \quad (11.13)$$

the rounding errors introduce a small multiple of the second solution

[†] E. T. Whittaker and G. N. Watson, *Modern Analysis* (C.U.P. 1927), ch. 16.

[‡] *Ibid.*, § 16.2.

$Y_n(x)$ of this recurrence relation. For $n < x$ this remains small, but for $n > x$ it behaves roughly as an increasing exponential, and increases without limit as n tends to infinity.

Thus this way of using the recurrence relation is not satisfactory for calculating Bessel functions $J_n(x)$ for $n > x$, though it is satisfactory for $n < x$. It would, however, be satisfactory for calculating $Y_n(x)$, since in this case the unwanted solution, of which a small multiple may be introduced by rounding errors, is one which decreases indefinitely, relative to the wanted solution $Y_n(x)$, as n increases.

On the other hand, the range over which $(n/x) > 1$ is just the range over which relaxation methods can be used effectively for the solution of (11.12), provided the solution has to satisfy two-point boundary conditions in n . This is the case for the Bessel function $J_n(x)$, since up to $n = x$ these can be built up satisfactorily by successive use of formula (11.13). This gives $J_n(n)$ as one terminal condition in n , and the other is given by $J_n(x) \rightarrow 0$ as $n \rightarrow \infty$.

This example shows how quite different results can be obtained by different ways of using the same simple formula; one way of using the recurrence relation (11.12) may lead to quite spurious results although no mistakes have been made in the calculation, whereas another way of using the same formula can be used to give results accurate to any assigned degree.

11.4. Smoothing

'Smoothness', either of a continuous function or of a set of discrete values, is a property of which it is difficult to give a quantitative definition. For a continuous function it implies smallness of high-order derivatives, and for a table of function values it implies smallness of the higher orders of differences; this implies also regularity of the differences, since if the n th differences are irregular, the $(n+10)$ th differences will not be small.

By 'smoothing' a set of function values is meant a process of replacing them by another set which differ only slightly from them but are 'smoother' in this sense. If each member of a set of function values has been obtained by an independent calculation, and each is subject to a rounding error, then the accuracy of the values may be increased somewhat by a smoothing process. But this improvement should not be relied on, and cannot be estimated. It should not be relied on because it is always possible that the rounding errors in a number of consecutive function values may be of the same sign and similar in magnitude, and

then smoothing will not improve them. Further, as we have already seen in Chapter IV, an incorrect set of function values may have smooth differences, and a set of correctly rounded-off function values may be less smooth than a set obtained by rounding off incorrectly. Also without knowing in some other way a more accurate set of function values, there is no criterion by which the improvement of the function values can be assessed; and if these more accurate function values were known, there would be no point in carrying out the smoothing process. If a set of function values is too much affected by rounding errors, the only reliable way of getting more accurate values is to carry out the calculation of the function values to greater numerical accuracy.

The main purpose in carrying out a process of smoothing must therefore be to achieve smoothness, not accuracy. The contexts in numerical analysis in which smoothness is a prime requirement are not many, so that such a process is not often required. But occasionally it is difficult to make satisfactory progress without one.

Consider, for example, the evaluation of a set of solutions of a differential equation involving a function $f(y)$ determined by experiment or by statistical sampling, the different solutions being distinguished by different initial conditions or different values of one or more parameters. For consistency between the various solutions, and also in order to use the differences of intermediate quantities for checking the integrations, it may be advisable to use in the numerical work a table of $f(y)$ which is smooth to a substantially greater degree of numerical accuracy than the accuracy of the experiments from which $f(y)$ is determined. This can sometimes be achieved by fitting an analytical formula to the experimentally determined values; when this has been done, the formula can be evaluated to any required numerical accuracy. But this process is inconvenient unless a relatively simple formula can be found to fit the experimental values within the experimental or sampling error, and it is also unnecessary. A more purely numerical smoothing process is often more useful and more effective.

Another example is provided by the process of § 10.72 for the integration of a parabolic partial differential equation. For one time interval of this process, as applied to the equation

$$\frac{\partial f}{\partial t} = \frac{\partial^2 f}{\partial x^2},$$

the equation (10.33) is solved with two-point boundary conditions in x . Suppose that in the process of evaluating the solution satisfying the

two-point boundary conditions, random rounding errors may occur in $[f(x, t + \delta t) + f(x, t)]$ up to $\pm q$ in the last significant figure kept. If the rounding error in $f(x, t)$ may be $\pm p$, that in $f(x, t + \delta t)$ may be $\pm(p + q)$, that in $f(x, t + 2\delta t)$ may be $\pm(p + 2q)$ and so on, and effects of rounding errors may be somewhat increased if Richardson's process of h^2 -extrapolation is used to correct approximately for the truncation error. Thus, however many figures are kept, the last will become more and more irregular as the calculation proceeds. If two or three guarding figures are kept, this will not affect the final results significantly, but such irregularities make it difficult to use differences for checking, and may lead to time being wasted in trying to find a suspected mistake that is not there. For this reason it is advisable occasionally to smooth $f(x, t)$ as a function of x during the progress of the calculation.

11.41. Automatic methods of smoothing

A simple example of one class of methods of smoothing is the following: Replace each function value f_j by the mean \bar{f}_j of five successive values of f centred on f_j , that is, take

$$\bar{f}_j = \frac{1}{5}(f_{j+2} + f_{j+1} + f_j + f_{j-1} + f_{j-2}).$$

This is sometimes called 'smoothing by fives', or 'smoothing by groups of five'. In this process, the irregularities get smoothed out by being distributed among neighbouring function values. This can be illustrated by the set of function values

$$f \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0, \quad (11.14)$$

for which this process gives

$$\bar{f} \quad 0 \quad 0 \quad \frac{1}{5} \quad \frac{1}{5} \quad \frac{1}{5} \quad \frac{1}{5} \quad \frac{1}{5} \quad 0 \quad 0. \quad (11.15)$$

The maximum value of $|\delta^2 \bar{f}|$ is $\frac{1}{5}$ whereas that of $|\delta^2 f|$ is 2: on the basis of this criterion, the set of values (11.15) is ten times as smooth as the set (11.14).

This process of smoothing by groups of $(2n + 1)$ can be repeated. For example, two successive processes of smoothing by groups of three, starting from the set of values (11.14) gives

$$\bar{f} \quad 0 \quad 0 \quad \frac{1}{9} \quad \frac{2}{9} \quad \frac{3}{9} \quad \frac{2}{9} \quad \frac{1}{9} \quad 0 \quad 0. \quad (11.16)$$

For this set of values, the maximum $|\delta^2 \bar{f}|$ is $\frac{2}{9}$.

These are two examples of a general method which consists of replacing each f_j by a linear combination

$$\bar{f}_j = \sum_{k=-n}^n a_k f_{j-k} \quad (11.17)$$

of function values centred on f_j . The smoothest set of function values

is simply $f_j = \text{constant}$, and in order that these should not be altered by the smoothing process, the coefficients must satisfy

$$\sum_{k=-n}^n a_k = 1;$$

and normally the coefficients will be symmetrical about $k = 0$. Different processes are given by different choices of the coefficients a_k in (11.17).

These methods are often unsatisfactory in practice for three reasons. First, once the particular smoothing formula to use has been decided, it is automatic in character in that the results are then determinate, and gives no opportunity for the exercise of judgement by the individual who is carrying out the calculation. This might at first sight seem an advantage, since the results will then be independent of the individual. But this apparent definiteness of the results is spurious since there is a good deal of latitude in the choice of what smoothing formula to adopt. And the smoothing process is in practice one in which it seems desirable to give the individual who is carrying it out some discretion on matters such as the degree of smoothing at which to aim and the degree to which changes $\bar{f}_j - f_j$ from the original function values are acceptable. Secondly, with methods depending on the use of formulae of the type (11.17), the smoothed values \bar{f}_j cover a smaller range of j than the original values; in a method due to Spencer,[†] recommended by Whittaker and Robinson,[‡] ten values at each end of the range are lost, so that from 30 values of f_j only 10 smoothed values in the middle of the range of j are obtained. Such a loss of range is often unacceptable. Thirdly, a special procedure is needed if it happens that some value of $f(x)$ is known exactly, such as a value $f(x) = 0$ at $x = 0$, and is not to be modified by the smoothing process.

The dangers of a blind use of an automatic smoothing process are illustrated in Fig. 30. Here the full curve is representative of the behaviour of the function $f(v) = R/v^2$, where R is the resistance of the air to a body moving through it at a speed of v ft./sec. If $f(v)$ is tabulated at intervals of 50 ft./sec. the behaviour of the second and higher differences of $f(v)$ is rather violent, and can be mollified by the application of a smoothing process. Spencer's process, applied to these data, gives results represented by the squares and broken curve in Fig. 30. They are certainly smoother (the greatest value of $|\delta^2 f|$ has been reduced from 232 to 48 in terms of the third decimal place as unit). But it does not follow that the smoothed values are a better representation of the

[†] J. Spencer, *J. Inst. Actuaries*, 38 (1904), 334.

[‡] *Calculus of Observations* (Blackie, 1940), p. 290.

actual behaviour of $f(v)$ than the unsmoothed values; they are almost certainly worse, and in particular the minimum about $v = 780$ ft./sec. is almost certainly spurious. But if one insists on using an automatic formula one has no control over the results it is going to give; if one

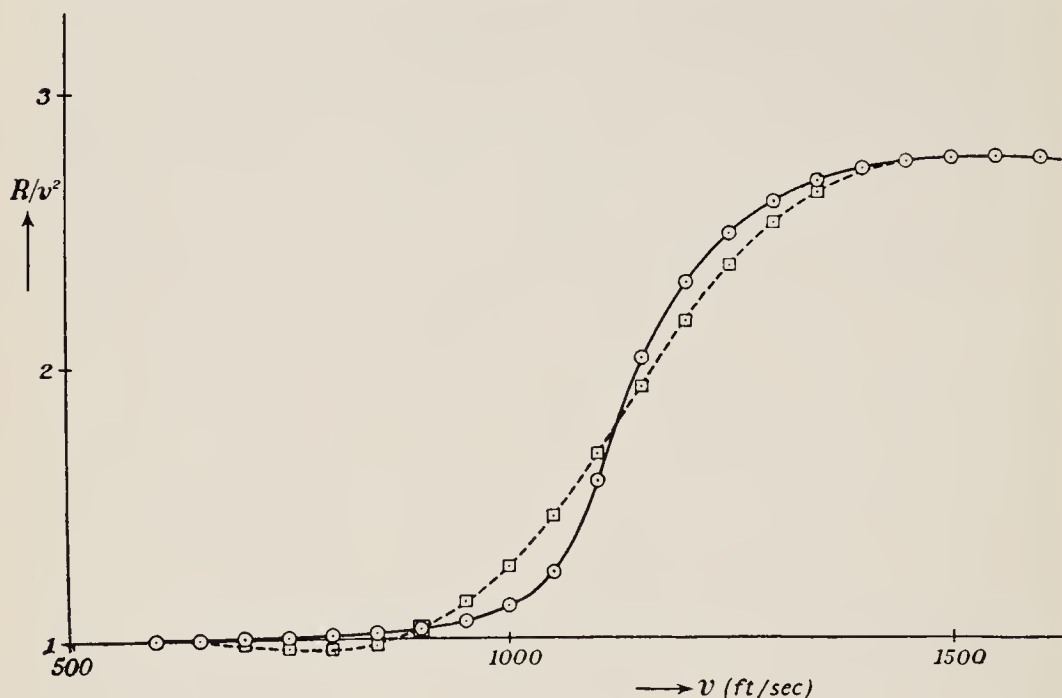


FIG. 30.

believes the use of the formula to be significant at all, all that one can do is to accept the results of using it.

11.42. Smoothing by use of an auxiliary function

A less formal but more practical method is due to A. T. Doodson.† It is based on the use of graphs.

Let $f(x)$ be the function which it is desired to smooth. Unless $f(x)$ is of only two or three figures, and sometimes even if it is of three figures, it will not generally be possible to smooth it directly by plotting and drawing a smooth curve 'through' the plotted points. But if $g(x)$ is a smooth function approximately equal to $f(x)$, it may be possible to plot the *difference* $f(x) - g(x)$ on a scale open enough to smooth it, to the degree of numerical accuracy required, by such a graphical process. Then $\bar{f}(x)$, the smoothed function by which $f(x)$ is replaced, is constructed as

$$\bar{f}(x) = g(x) + \text{smoothed}\{f(x) - g(x)\}.$$

† This method was devised during the war of 1914–18, in connexion with ballistic work, but only published recently, in *Quart. J. Mech. and Applied Math.* 3 (1950), 217.

The auxiliary function $g(x)$ can be formed in several ways. It may, for example, be taken to be given by an analytical formula, such as ax^2 , $ax/(x^2+b^2)$, e^{ax} , be^{-ax^2} , if there is any theoretical reason or empirical indication that $f(x)$ is approximately of such a form. Another process is to build up $g(x)$ from a smooth set of differences. This is of more practical use in many cases, as it can be used equally well whether or not a good approximation to $f(x)$ can be obtained by a simple analytical formula, and it does not involve any selection and adjustment of parameters in an analytical formula so as to get a good overall fit to $f(x)$.

For simplicity suppose $f(x)$ to be given at equal intervals of x . And as an example of the general process, suppose this function, and the interval of tabulation, to be such that the range of the values of $\delta^2 f$ is not more than 200, so that they can be plotted on such a scale (1 mm. or $\frac{1}{20}$ in. to a unit) that they can be read off to a unit.

The process is then as follows. Plot $\delta^2 f(x)$ and draw 'through' the plotted points as smooth a curve as possible without smoothing away significant features of the behaviour of the second differences. This is one place at which discretion is required in judging what features of the behaviour of the second differences are significant. If the uncertainty of each value of $f(x)$ is known, the range of uncertainty of each second difference can be found and indicated on the plot, and this may help in distinguishing significant from non-significant features of the variation of the second differences. It is better to over-smooth at this stage rather than the reverse; significant variations which are smoothed out at this stage are replaced at a later stage.

Let $h(x)$ be these smoothed values of $\delta^2 f(x)$. It is not advisable to double-sum them directly to give the auxiliary function $g(x)$. A small systematic difference between two different ways of drawing the curve from which $h(x)$ is read off may build up, on double-summing, to a substantial amount, so that though one curve might give a $g(x)$ which differed little from $f(x)$, the other might give a $g(x)$ departing from it to such an extent that it would be difficult to plot $f(x) - g(x)$ on an adequate scale; and there can be no certainty that the curve from which $h(x)$ is read off is not of the latter kind. It is therefore best first to form the single sum $\sigma h(x)$ of the values of $h(x)$, and to modify this if necessary so as to get a general agreement with the first differences of $f(x)$, before forming an auxiliary function. The differences

$$\delta f(x) - \sigma h(x)$$

between the first differences of $f(x)$ and the first sum of $h(x)$ are therefore

plotted and smoothed graphically. The smoothed values of

$$\delta f(x) - \sigma h(x)$$

are then added to the values of $\sigma h(x)$ to give the first differences of the auxiliary function $g(x)$, which is then built up from these differences. Thus $g(x)$ is given by

$$g = \sigma[\sigma h + \text{smoothed}(\delta f - \sigma h)].$$

Finally, $(f-g)$ is plotted and smoothed, and the smoothed function $\bar{f}(x)$ is given by

$$\bar{f}(x) = g(x) + \text{smoothed}\{f(x) - g(x)\}.$$

In this final stage discretion can again be exercised regarding the extent to which values of f may be modified by the smoothing process, and the significance of various features of the behaviour of f . The difference between the original and smoothed values of f at any value of x is

$$f(x) - \bar{f}(x) = \{f(x) - g(x)\} - \text{smoothed}\{f(x) - g(x)\},$$

and the right-hand side here is the departure of the smooth curve from the plotted point $(f-g)$ at each value of x . If the range of uncertainty of each value of f , or the maximum change in each value which would be acceptable, is known, this can be indicated on the plot, and the smooth curve drawn so that its departures from the plotted points do not exceed this range at any point. In particular, if at any point the value of $f(x)$ is known exactly, the curve of $(f-g)$ must be drawn to pass through the plotted point at that value of x . In Doodson's method of smoothing, particular features such as this can be taken into account quite easily and without departure from the regular procedure.

The differences of the final values of $\bar{f}(x)$ provide a check of the calculation and an indication of the degree of smoothness which has been obtained. An indication of the extent to which the final results depend on the details of the process of smoothing is given by carrying out the process twice using the same set of values of $f(x)$ but different smooth curves from which to read off the smoothed second differences of $h(x)$.

The process can be adapted to start from other orders of differences than the second. However, if $f(x)$ is at all seriously irregular, the higher differences of $f(x)$ probably vary so wildly that it is difficult to see any general trend in their values.

Examples of this process, and its extension to functions of two variables, can be found in Doodson's paper.

XII

ORGANIZATION OF CALCULATIONS FOR AN AUTOMATIC MACHINE

12.1. Automatic digital calculating machines

WHEN we write a number in the ordinary way, such as 1925, the symbols such as 1, 9, 2, 5 in this example stand for what we call the *digits* of a number, and a piece of equipment which operates directly with, and records, the discrete digits of each number is often called a *digital calculating machine*. Since about 1938 there has been a great development of such machines with two important features. First, they can carry out long and intricate numerical calculations quite automatically once they have been provided with a specification, in a suitable form, of the calculation to be carried out. And, secondly, they are very versatile, so that the same machine can be used for many quite different kinds of calculation; for example, for calculating values of a function from its power-series expansion, for solving large systems of linear simultaneous equations, for finding the characteristic values of matrices, and for the step-by-step integration of ordinary differential equations. To express these two features, such machines are sometimes called *general-purpose, automatic, digital calculating machines*.

The process of organizing calculations for such machines is a branch of numerical analysis which has only come into being with the machines, and this chapter is included here to give an introduction to the subject. It is concerned with the planning of calculations for such machines rather than with the machines themselves; it is only concerned with the machines in so far as their characteristics affect the process of organizing calculations for them.

To see what is required of such a machine, consider first the organization of a calculation carried out by hand with the assistance of a desk machine. This is represented diagrammatically in Fig. 31. There are three kinds of equipment the computer has to assist him; these are represented by rectangular blocks in the figure. One is a desk machine, another is a set of tables, and the third is the working sheet on which intermediate and final results will be recorded and on which should be written enough data to identify the calculation and to summarize the calculating procedure.

When the method for doing a calculation has been decided, the detailed

process of carrying it out consists of (i) a sequence of arithmetical operations carried out on the machine, or perhaps on a slide-rule as auxiliary equipment, or mentally in the case of simple operations such as multiplication or division by 2 or addition of pairs of numbers, and (ii) transfer of numbers between the three blocks represented in Fig. 31.

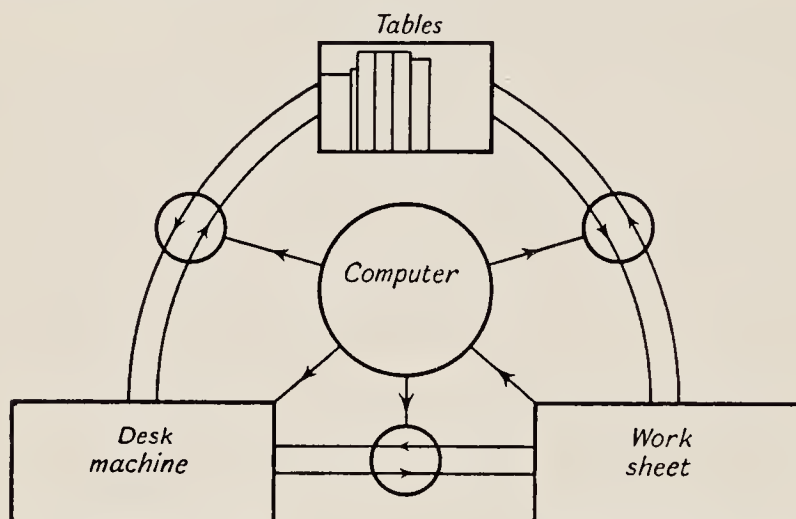


FIG. 31.

The transfers and the arithmetical operations are controlled by the individual who is carrying out the calculation, who is represented by the large circle in Fig. 31; the controls he exerts are represented by directed lines from the controller to small circles representing control of the transfer of numbers, and to the desk machine. He also takes from the working sheet information about the arithmetical operations to be carried out and their sequence; this is represented by the directed line from the work sheet to the computer.

An automatic machine must be capable of carrying out the same processes, and can be thought of as having a similar organization, as shown diagrammatically in Fig. 32. It must have an *arithmetical unit* in which arithmetical operations can be carried out, to take the place of the desk machine in a hand calculation; a *store* both for numbers and for operating instructions, to take the place of the work sheet and tables; and a *control system* to take the place of the human computer who controls the sequence of operations in a hand calculation. The machine also needs input and output equipment for receiving numerical data and operating instructions from the outside world and for delivering its results.

Whatever the physical form of the store, it must provide a number of identifiable *storage locations*, and it is convenient to think of these as

distinguished by being numbered. The number which is the label of any storage location is often called its 'address', or the 'address' of its content. $C(n)$ will be used for 'the content of storage location n '. It is sometimes convenient to represent the address of the number which is the value of a quantity x by $L(x)$ or $A(x)$, or by $C^{-1}(x)$ if it is desired to emphasize that the relation $n = L(x)$ is the inverse of the relation $x = C(n)$.

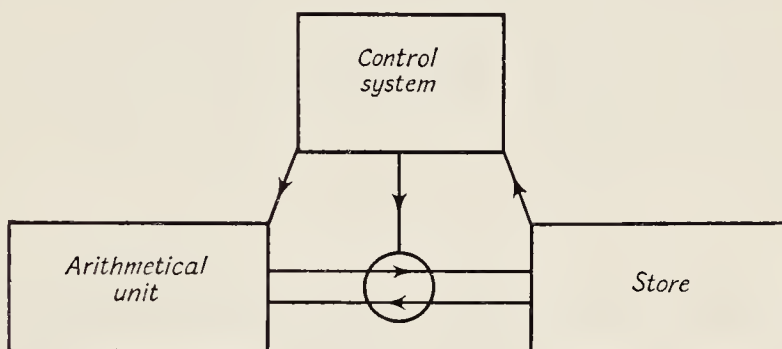


FIG. 32.

The specification of an operation which the machine is required to carry out is called an 'instruction' or 'order' (sometimes a 'command'), and the ordered set of such instructions needed to carry out a calculation is called the 'program' for that calculation. An important feature of most machines is that instructions are coded in such a way that they have the same form, within the machine, as numbers, the difference between numbers and instructions being in the way they are used. The content of a storage location is then usually called a 'word', whether it represents a number or an instruction; 'words' representing numbers are normally used by being transferred between the store and the arithmetical unit, and 'words' representing instructions are normally used by being transferred to the control system. But since there is no distinction, within the machine, between numbers and instructions, it is possible to use the arithmetical unit to build up, transfer, and modify the operating instructions themselves as the calculation proceeds.

The effect of this on the organization of a calculation is so profound that there may be little relation between the processes of organizing a calculation for machines which do and for those which do not provide this possibility. However, its importance is now well realized, and it is provided on most of the machines already (1957) in operation and is likely to be provided on all future machines. It will be assumed in the rest of this chapter.

There are two main forms for instructions; these can be illustrated by an example. Suppose we want the machine to form the sum of the contents of storage locations n_1 and n_2 and to put the result into location n_3 . This could be done by a single instruction which could be written symbolically

$$C(n_1) + C(n_2) \text{ to } n_3. \quad (12.1)$$

An alternative is as follows. Suppose that the arithmetical unit is of a kind which includes as one of its components a register, usually called an 'accumulator', corresponding to the 'product register' or 'accumulator' of a desk machine, which accumulates the sum of numbers added into it until it is cleared; the content of the accumulator will be written $C(Acc)$. Then the required operation can be done by the three separate instructions:

$$C(n_1) \text{ to } Acc, \quad C(n_2) \text{ to } Acc, \quad C(Acc) \text{ to } n_3. \quad (12.2)$$

Each of the instructions of the form (12.1) specifies three addresses in the store, whereas each instruction of the form (12.2) specifies a single such address. These forms of instruction are consequently known as the 'three-address' and 'one-address' forms respectively.

An instruction specifies an operation to be carried out. But it is also necessary to specify the sequence in which such operations are to be carried out; that is to say, after carrying out one operation, the machine must be enabled to select the instruction for the next operation. In a machine in which instructions are contained in the same store as numbers this can be done in two ways. One is to include in each instruction the address from which the *next* instruction is to be taken; with a three-address specification of the operation to be carried out, this gives altogether a 'four-address' form of instruction. Another way is normally to store instructions at addresses numbered serially in the same order as the time-sequence in which they are to be carried out. This will be referred to as 'serial storage' of instructions. In this case the address of the instruction currently being carried out is recorded in a register whose content is normally increased by unity on the completion of this instruction, and the content of this register is used to control the selection of the next instruction. Then an explicit specification of the address of the next instruction is needed only when it is required to depart from the serial order in which the instructions are located in the store.

The control system of a machine will depend on the standard form adopted for instructions and the means adopted for selecting the next instruction, and once the machine is built, the instructions used must conform to the type for which the control system has been designed.

These features, rather than the physical form of the store or of the arithmetical unit, are the essential features of a machine from the point of view of the user.

12.2. Preparation of calculations for an automatic digital calculating machine

The process of preparing a calculation for an automatic digital machine can be broken down into two parts, often called 'programming' and 'coding'.

By the 'program' for a calculation is meant the schedule of operating instructions which has to be provided to the machine in order that it shall carry out the calculation. 'Programming' is the process of planning the sequence of operating instructions required, and 'coding' is the process of translating these instructions into the particular form in which they are supplied to the machine. In simple calculations these are hardly two distinct processes, but in more elaborate calculations it is convenient to treat them as separate.

A process of programming is required in a hand calculation; before we can start doing any calculation we must decide just how we are going to do it. For work with an automatic machine, programming may involve breaking down the calculation to a sequence of the elementary operations, such as addition, multiplication, and selection of the next instruction, which the machine can carry out. But the machine and the process of providing it with instructions may be such that groups of operations for standard processes, such as evaluation of $\cos x$ given

the value of x , or of $\int_a^b f(x) dx$ given the values of $f(x)$ at a set of values

of x , can be programmed and coded once for all. If this can be done, each such process can be regarded as a unit in programming a calculation, and not analysed into elementary operations. The program for such a standard process will be called a 'sub-routine'. The possibility of using such sub-routines freely greatly lightens the work of preparing a calculation for a machine, and the machine and the form of its instructions should be planned to provide this facility. The use of the same form in the machine for instructions and for numbers, and the freedom which this gives to modify instructions by arithmetical and other operations on them, are important features in making it easy to provide and exploit the possibilities of using such sub-routines. To a potential user of an automatic machine, means of organizing calculations for it are as important as the provision of the machine itself, and the provision of a library

of sub-routines for standard processes is an important step in this direction.

Although the various kinds of machines differ considerably from one another in their internal organization and operation, the general process of programming a calculation will be much the same for any of them, for it depends primarily on the structure of the sequence of operating instructions required to carry out the calculation. Some characteristic features on an individual machine may, however, affect the details of the programming. Such features are

- (i) The standard form of operating instructions adopted; whether this is, for example, a one-address or four-address form.
- (ii) The facilities provided by the standard instructions; for example, whether division can be carried out directly or has to be done by means of an iterative process which has to be programmed.
- (iii) The criteria which it is possible to use for discrimination between possible alternative courses of procedure; for example, whether it is only possible to discriminate on the sign of a number or also on criteria such as the likeness or unlikeness of signs of two numbers.

The process of coding does, however, depend on the particular machine. It does not depend primarily on the physical form of the store or on the way in which numbers are represented in the machine, but on two features, namely the standard form of instructions and the means of selecting the next instruction, as explained at the end of § 12.1.

The details of programming and coding a calculation will be different for different machines, depending on the facilities provided by the machine itself and by the organization associated with it (such as the range of the available library of sub-routines).†

12.3. Hand and automatic calculation

Almost any method for doing a calculation by hand, that is, with a desk machine but without the use of an automatic machine, can be

† A fairly full account of programming and coding for one machine, the EDSAC at the Mathematical Laboratory at the University of Cambridge in its original form, is given in M. V. Wilkes, D. J. Wheeler, and S. Gill, *The Preparation of Programs for an Electronic Digital Calculating Machine* (Addison-Wesley Press, Cambridge, Mass., 1951; a second edition is in course of preparation). See also A. D. and K. H. V. Booth, *Automatic Digital Computers* (Butterworth, 2nd edn., 1956), chapters 13–16; W. J. Eckert and R. James, *Faster, faster* (an account of the U.S. Naval Ordnance Research Computer NORC and methods of programming for it (McGraw-Hill, 1955)); M. V. Wilkes, *Automatic Digital Computers* (Methuen, 1956), ch. 3; J. H. Wilkinson, *Phil. Trans. Roy. Soc.*, **248** (1955), 243; and a number of shorter articles in *M.T.A.C.*

programmed for an automatic machine. A possible exception is the relaxation process (§§ 8.5, 10.61) for which it would be difficult to formalize the judgements one uses in practice about when and by how much to over-relax, and when to use group relaxations, and to express these judgements in terms of operating instructions to an automatic machine. But it does not follow that the best method for a hand calculation is the best method for an automatic machine. There are three main reasons for this difference.

First, in most hand calculations of any magnitude, the time taken in carrying out the numerical work is substantially longer than the time taken in planning it, whereas with an automatic machine the time taken to carry it out may be shorter than the time taken to program and code it. Thus in a hand calculation it is worth spending some time in planning the calculation to save numerical work, whereas with an automatic machine it may be best to obtain the same results by a simple process involving a large number of steps to save the time that would be taken in planning, programming, and coding a less simple method using fewer numerical steps. For example, on an automatic machine a relatively large number of repetitions of a simple first-order iterative process (§ 9.3) may be preferable to a smaller number of repetitions of a more complicated second-order process. And in calculating an integral as a function of the upper limit, it might be best with an automatic machine to use a very simple integration formula, such as Simpson's rule or even the trapezium rule, with a large number of short intervals, 0.01, or 0.005, or even perhaps 0.001, when in a hand calculation one might prefer to use an integration formula to sixth or eighth differences of the integrand, with interval 0.1.

Secondly, the storage capacity of a machine is limited, whereas that of the working sheets of a hand calculation is practically unlimited. This has several reactions on programming for an automatic machine. For example: (i) use of many repetitions of a simple procedure which can be programmed in a few instructions is preferable to a few repetitions of a more elaborate procedure for which the longer program would take more storage space; (ii) a strictly repetitive procedure is to be preferred to a procedure which is mainly repetitive but for which special occasional processes have to be used in addition: for example, in some processes for the numerical integration of differential equations a special procedure is needed for the first interval of the integration; the instructions for this special procedure will take up some storage space but will be used once only for each solution, and a method which does not require a special

starting process may be preferred; (iii) it may be preferable to calculate values of standard functions, such as circular and exponential functions and their inverses, as they are required, rather than to store tables and the instructions for entering them and interpolating in them.

And thirdly it is usually no shorter or easier to calculate with simple numbers than with numbers of many digits. For example, if e^y is calculated from a series, then in calculating, say, $\int_0^1 e^{x^2} dx$ by a Gauss formula (see § 6.61) the fact that e^{x^2} is required for values of x such as 0.230765 (x_2 for a five-point Gauss formula, see § 6.61) and not only for simple values is no drawback.

All these differences have considerable influence on the choice of methods for carrying out calculations by automatic machines. For example, for evaluating an integral between fixed limits formulae of the Gauss type are much more attractive for work with an automatic machine than for a hand computation. Also for evaluating an integral as a function of its upper limit it might even be better to do a number of independent integrations by means of a Gauss formula, with different values of the upper limit, rather than to build up the integral by accumulating a sequence of contributions to it. And for the solution of partial differential equations of elliptic type, a form of the Richardson-Liebmann process (§ 10.64) may be more convenient for an automatic machine than the relaxation process.

Some work has been done on the development of methods particularly suited to the capabilities and limitations of automatic machines, but the main developments of this branch of numerical analysis probably still lie in the future.

EXAMPLES

Note: Several of the following examples are specimens of types of which the reader can make up other examples for himself. For instance, in the first example $e^{0.1}$ could be replaced by some other number, and in Example 6, the series to be evaluated could be replaced by the series solution of some other second-order linear differential equation with the first derivative absent.

1. Given $e^{0.1} = 1.105171$ to six decimals:

- (i) Calculate $y_n = e^{0.1n}$ up to $n = 10$ by successive multiplication and transfer.
- (ii) Check the results by verifying the following relations between the y_n 's:

$$\begin{array}{ll} y_6 \times y_4 = y_{10}, & y_{10}/y_7 = y_3, \\ y_7 \times y_2 = y_9, & y_9/y_5 = y_4, \\ y_5 \times y_3 = y_8, & y_8/y_6 = y_2. \end{array}$$

- (iii) Check the results by differencing the values of y_n (including the value $y_0 = 1$) to second differences, and verifying that $(\delta^2 y_n)/y_n$ is constant.

(*Note:* The main purpose of this example is to give practice in the use of a desk machine; the method of checking in section (ii) of the example is not recommended as a standard procedure for regular use.)

2. Given $e^{0.125} = 1.133148$, evaluate $2[(\cosh 0.125) - 1]$ without writing down any intermediate results.

Prove the relation $\delta^2 e^x = 2[(\cosh \delta x) - 1]e^x$ and use it to build up $e^{0.125n}$ up to $n = 10$.

3. Given $\sin 10^\circ = 0.1736482$, find $(1 - \cos 10^\circ)$ to seven decimals by an iterative process based on the formula

$$1 - \cos x = \sin^2 x / [2 - (1 - \cos x)].$$

Prove the relation $\delta^2(\sin x) = -2(1 - \cos \delta x)\sin x$, and use it to build up a table of $\sin(n \cdot 10^\circ)$ as far as $n = 9$.

4. Calculate $\sum_n x_n / \sum_n x_n^2$ on a desk machine for $x_1 = 1.274, x_2 = 0.984, x_3 = 1.577, x_4 = 0.126$ without writing down any intermediate results.

5. Find $3 \times (\text{£}3.8s.7d.) + 19 \times (\text{£}1.17s.9d.) + 16 \times (\text{£}2.2s.11d.)$ using an ordinary desk machine, setting the sterling amounts in £.s.d. and exhibiting the results in £.s.d. , without writing down any intermediate results.

(*Note:* Assign the three right-hand places on the setting levers or keyboard to pence, the next three to shillings, and the rest to pounds. After forming the sum, reduce the number of pence by a multiple of 12, and add that same multiple to the shillings, by adding 988 in the right-hand three places until the number of pence in the result is less than 12. Treat the shillings similarly.)†

6. Evaluate
$$y = \frac{1}{2}x^2 + \frac{1}{2 \cdot 4 \cdot 5}x^5 + \frac{1}{2 \cdot 4 \cdot 5 \cdot 7 \cdot 8}x^8 + \dots$$

to five decimals for $x = 0(0.2)2.0$, keeping seven decimals in the individual terms and rounding off the sums to five decimals.

† This procedure for using a decimal machine for certain calculations in sterling was shown to me by Dr. L. J. Comrie.

Check the results by evaluating y'' from the differential equation $y'' = 1 + xy$ satisfied by this function y , and verifying the relation

$$\delta^2 y_j = (\delta x)^2 [y_j'' + \frac{1}{12} \delta^2 y_j'' - \frac{1}{240} \delta^4 y_j''] + O(\delta x)^8.$$

7. Show that the function $y = (\sin x - x \cos x)/x$ satisfies the differential equation $y'' + (1 - 2/x^2)y = 0$.

Evaluate this function to five decimals, for $x = 0(0.1)2.2$, from its power series expansion, and check the results by use of the differential equation.

8. It is given that if $l(a, b)$ is the common limit of the sequences $\{a_n\}, \{b_n\}$ given by

$$\left. \begin{aligned} a_0 &= a, & a_{n+1} &= \frac{1}{2}(a_n + b_n) \\ b_0 &= b, & b_{n+1} &= \sqrt{a_n b_n} \end{aligned} \right\} \quad (n \geq 0)$$

then†
$$\int_0^{\frac{1}{2}\pi} d\theta / (a^2 \cos^2 \theta + b^2 \sin^2 \theta)^{\frac{1}{2}} = \pi / 2l(a, b). \quad (\text{A})$$

It is also given that the complete elliptic integrals

$$K(k) = \int_0^{\frac{1}{2}\pi} d\theta / (1 - k^2 \sin^2 \theta)^{\frac{1}{2}}, \quad E(k) = \int_0^{\frac{1}{2}\pi} (1 - k^2 \sin^2 \theta)^{\frac{1}{2}} d\theta$$

satisfy the differential equations

$$dK/dk = [E - (1 - k^2)K]/k(1 - k^2), \quad dE/dk = (E - K)/k. \quad (\text{B})$$

Use the result (A) to evaluate $(2/\pi)K(k)$ to five decimals for $k^2 = 0(0.1)0.8$. Show from equations (B) that $K(k)$, as a function of k^2 , satisfies the equation

$$\frac{d}{d(k^2)} \left[k^2(1 - k^2) \frac{dK}{d(k^2)} \right] = \frac{1}{4}K$$

and use this equation to check the values of $K(k)$ derived by use of formula (A).

9. Build up the cubic $f(x) = x^3 - 5x^2 + 6x + 1$ between $x = 2$ and $x = 3$ at intervals of 0.1 by means of a difference table. From these results estimate the position x_m and magnitude of the minimum of $f(x)$ near $x = 2.5$. Verify by solving the quadratic for x_m and evaluating $f(x_m)$.

(Note: Change to $\xi = x - 2$ as variable, verifying the transformed form of the cubic by evaluating it for $\xi = 0, \pm 1, \pm 2$ and comparing with the values of $f(x)$ on p. 42.)

10. Evaluate $0.623x^3 - 1.876x^2 + 5.623x + 2.875$ to three decimals for

$$x = 0(0.32)2.56$$

and check the results by differencing.

11. The following values are alleged to be copied from a table of $x^{\frac{1}{2}}$. Locate and correct the mistakes by examination of the differences.

| x | $f(x)$ | x | $f(x)$ |
|-----|---------|-----|---------|
| 27 | 3.00000 | 35 | 3.27107 |
| 28 | .03659 | 36 | .30193 |
| 29 | .07232 | 37 | .33332 |
| 30 | .10723 | 38 | .36198 |
| 31 | .14318 | 39 | .39121 |
| 32 | .17480 | 40 | .41995 |
| 33 | .20753 | 41 | .44852 |
| 34 | 3.23961 | 42 | 3.47603 |

† See E. T. Whittaker and G. N. Watson, *Modern Analysis*, ch. 22, example 46.

12. Using 6-figure tables of $\sin x^\circ$, calculate the function

$$y = \sin x^\circ - 2 \cdot 10^{-4} \left[\frac{\sin(x-50)\pi/10}{(x-50)\pi/10} \right] \exp[-(x-50)^2/100]$$

for $x = 30(1)70$, and round off to five decimals.

Compare the second and fourth differences of the rounded values of y with those of five-figure values of $\sin x^\circ$.

Repeat for $x = 25(5)75$ and for $x = -20(10)120$.

(Note: $y - \sin x^\circ$ can be regarded as an 'error' in a table of $\sin x^\circ$. The purpose of this example is to illustrate that smooth differences do not necessarily imply freedom from error; the differences of y at intervals 1 in x are no more irregular than those of $\sin x$. It also shows that a table may appear smooth on a small scale as represented by the differences at a small interval of x , but unsmooth on a large scale.)

13. Show that the sum $\sum_{j=0}^{J-1} \delta^2 f_{2j+1}$, of alternate second differences, can be expressed in terms of the operator $U = (\delta x) d/dx$ as $(\tanh \frac{1}{2} U)(f_{2J} - f_0)$.

Deduce the value of this sum when f is a periodic function which is an even function both of $(x - x_0)$ and of $(x - x_{2J})$, and examine whether or not this result is independent of rounding errors in the f values.

14. From the table of the function y calculated in Example 6:

- (i) Use the 'half-way' interpolation formula to obtain y at $x = 0.7, 0.9, 1.1, 1.3$.
- (ii) Interpolate y for $x = 0.95(0.01)1.00$ by Everett's formula.
- (iii) Find the value of x for which $y = 0.5$
 - (a) by inverse interpolation using the values at 0.1 intervals only;
 - (b) by inverse interpolation using the values at 0.01 intervals calculated under (ii).

15. From a table of $\sin x$ at intervals of 10° in x (see p. 60):

- (i) Find $\sin 23^\circ 20'$ and $\sin 26^\circ 40'$.
- (ii) Find $\sin^{-1} 0.40$

- (a) by inverse interpolation using a formula involving the differences of $\sin x$ as a function of x ;
- (b) by using Lagrange's interpolation formula, treating x as a function of $\sin x$, and verifying by interpolating in the table of $\sin x$ for the value of $\sin^{-1} 0.40$ obtained;
- (c) by using the divided differences of x as a function of $\sin x$.

16. Construct a table of values of $\log(n!)$ to five decimals for $n = 5(1)12$. Use this table to interpolate $\log(x!)$ for $x = 8\frac{1}{2}$ and $9\frac{1}{2}$, and verify that the interpolated values satisfy the relation $(9\frac{1}{2})!/(8\frac{1}{2})! = 9\frac{1}{2}$.

Derive a value for $(\frac{1}{2})! = \frac{1}{2}\sqrt{\pi}$, and hence a value of π .

17. Given the values

| | | | | | |
|---------|---|---|----|----|-----|
| $x = 0$ | 1 | 2 | 3 | 4 | 5 |
| $y = 0$ | 1 | 8 | 27 | 64 | 125 |

examine the result of attempting to interpolate x for $y = 20$ by a six-point Lagrange formula for x as a function of y .

If the calculation were done by the use of divided differences what symptoms would suggest that the result should be accepted with suspicion?

18. Continue to $x = 0.285$ the subtabulation started in the example in § 5.61.

19. Continue to $x = 1.6$, by intervals $x = 0.05$, the evaluation of $\int_0^x e^{w^2} dw$ started in the example in § 6.4.

20. Show that

$$\int_{x-1}^{x_2} f(x) dx = 4(\delta x) \left[1 + \frac{2}{3}\delta^2 + \frac{7}{90}\delta^4 - \frac{2}{945}\delta^6 + \frac{13}{56700}\delta^8 \right] f_0 + O(\delta x)^{11}.$$

21. The function $\operatorname{erfc} x$ is defined by

$$\operatorname{erfc} x = (2/\pi^{\frac{1}{2}}) \int_x^{\infty} e^{-w^2} dw.$$

Evaluate $\operatorname{erfc} x$ for $x = 0(0.1)1.2$ to five decimals by quadrature. Evaluate

$2 \int_x^{\infty} \operatorname{erfc} w dw$ for $x = 0(0.1)1.0$ by quadrature and check by use of the relation

$$2 \int_x^{\infty} \operatorname{erfc} w dw = (2/\pi^{\frac{1}{2}}) e^{-x^2} - 2x \operatorname{erfc} x.$$

(Notes: (i) $\operatorname{erfc} 0 = 1$; $2 \int_0^{\infty} \operatorname{erfc} w dw = 2/\pi^{\frac{1}{2}} = 1.28379$; (ii) the relation between

$\int_x^{\infty} \operatorname{erfc} w dw$ and $\operatorname{erfc} x$ is obtained by integrating by parts.)

22. Evaluate $\int_0^{\infty} [e^{-x^2}/(x+1)] dx$ and $\int_0^{\infty} [e^{-x^2}/(x+2)] dx$ to five decimals by quadrature. Check by evaluating the difference $\int_0^{\infty} [e^{-x^2}/(x+1)(x+2)] dx$ between these integrals by an independent quadrature.

23. Evaluate $y = (1/\pi) \int_0^{\pi} \cos(x \sin \theta) d\theta$, to four decimals, by quadrature for $x = 0(\frac{1}{4}\pi)\frac{5}{4}\pi$. Estimate, as closely as you can from the results, the smallest positive value of x for which $y = 0$.

24. Continue to $x = 1.6$ the integration of the equation $y'' = (1-x^2)y$ started in the worked example in § 7.2.

25. Continue to $x = 2.0$ the integration of the equation $y' = 1 - 2xy$ started in the example in § 7.3. The solution of this equation is $y = e^{-x^2} \int_0^x e^{w^2} dw$. Compare the results of the integration of this differential equation with those of the worked example in § 6.4 and its continuation in Example 19.

26. The function $f(x) = \int_0^{\infty} e^{-u^2}/(x+u) du$ satisfies the differential equation

$$f' + 2xf = -\frac{1}{x} + \pi^{\frac{1}{2}} \quad (\text{see § 6.56}).$$

Starting from the value of $f(1)$ obtained in Example 22, integrate this equation as far as $x = 2$. Compare the value of $f(2)$ obtained by integration with the value of $f(2)$ obtained in Example 22.

27. Evaluate $y = (3/\pi) \int_0^{1/\pi} e^{ix \cos \theta} d\theta$ to four decimals for $x = 0(\frac{1}{2})2$

(a) by expanding the integrand in series, integrating term by term with respect to θ , and evaluating the resulting series in x ;

(b) by quadrature;

(c) by obtaining a second-order differential equation satisfied by y as a function of x , and evaluating the appropriate solution by numerical integration.

28. Find, to two decimals, the solution of the equations

$$18x - 4y + 3z = 53,$$

$$10x + 16y + 2z = 87,$$

$$5x + 3y + 9z = 21$$

(a) by elimination; (b) by relaxation.

29. A cubic $y = a_0 x^3 + a_1 x^2 + a_2 x + a_3$ takes the values $y = 12, 6, 0, 12$ for $x = -2, 0, 1, 3$ respectively. Find the values of the coefficients a_0, a_1, a_2, a_3

(i) by substituting $x = -2, 0, 1, 3$ and solving the resulting simultaneous equations for the coefficients;

(ii) by use of divided differences (see § 5.72).

30. Use Milne's method (§ 6.8) to obtain an expression for the error term of the integration formula

$$\int_{x_0}^{x_1} f(x) dx = \frac{1}{2}(\delta x)[f_0 + f_1 - \frac{1}{6}(\delta x)(f'_1 - f'_0)].$$

31. (i) Invert the matrix

$$\begin{pmatrix} 5 & 7 & 6 & 5 \\ 7 & 10 & 8 & 7 \\ 6 & 8 & 10 & 9 \\ 5 & 7 & 9 & 10 \end{pmatrix}.$$

(ii) Find the characteristic values and characteristic vectors of this matrix.

32. Construct the inverse of the matrix

$$\begin{pmatrix} -23 & 11 & 1 \\ 11 & -3 & 2 \\ 1 & 2 & 1 \end{pmatrix}$$

from its characteristic vectors and the reciprocals of its characteristic values as determined in § 8.72.

(Note: This is the matrix whose inverse is found by elimination in the worked example in § 8.3.)

33. Solve the equations

$$14x_1 + 7x_2 + 17x_3 + 8x_4 = 134,$$

$$7x_1 + 11x_2 + 13x_3 + 4x_4 = 70,$$

$$17x_1 + 13x_2 + 42x_3 - 11x_4 = 77,$$

$$8x_1 + 4x_2 - 11x_3 + 30x_4 = 70.$$

34. Find the solution of the equation $y'' = x^2y - 1$ for which $y = 0$ at $x = \pm 2$
- by evaluating a particular integral and a complementary function by step-by-step integration and forming the appropriate linear combination;
 - by a relaxation method.

35. Work out a relaxation method for finding a solution of $\nabla^2 V = 0$ for a system with symmetry about an axis. Apply it to find an approximate solution of $\nabla^2 V = 0$ for the axially-symmetrical system formed by rotating Fig. 20 (p. 246) about its axis of symmetry.

36. $\nabla^2 V = 0$ in the space between the planes $z = 0, a$. On the plane $z = 0$, $V = 0$ for $r > a$ and $V = J_0(kr)$ for $r < a$, where ka is the first root of $J_0(x) = 0$; on the plane $z = 2a$, $V = 0$. Find to two decimals the variation of V in the space between the planes.

37. V satisfies the equation $\nabla^2 V = 2/a^2$ inside a square of side a , and $V = 0$ on the boundary of the square. Find to three decimals the value of V at the centre of the square.

38. V satisfies the equation $\nabla^2 V = -320/a^2$ within a circular cylinder of radius a and length $2a$, and $V = 0$ on the bounding surface. If (R, z) are radial and axial coordinates in a section containing the axis of the cylinder, with origin at its centre, obtain the finite difference equations for V on a grid of square mesh with $\delta R = \delta z = a/n$ (n integral), and solve them to two decimals for $n = 2$ and 4 . Use the results to estimate, by Richardson's process (§ 7.51), the value of V at the centre of the cylinder for the solution of the differential equation.

(Note: For $n = 2$, use a direct method for the solution of the finite-difference equations; for $n = 4$, use a relaxation procedure.)

39. Find to three decimals the coefficients in the quadratic factors of

$$f(z) \equiv z^4 - 4.0z^3 + 7.8z^2 - 8.2z + 5.6.$$

Hence obtain to three decimals the roots of $f(x) = 0$.

40. Find to two decimals the other roots of the simultaneous non-linear equations of which one root is found in § 9.6.

41. Use the iterative formulae

$$(a) \quad y_{n+1} = \frac{1}{2}[y_n + (a/y_n)], \quad (b) \quad y_{n+1} = y_n(3a - y_n^2)/2a$$

for $a^{\frac{1}{2}}$ to evaluate $\sqrt{5}$ and $\sqrt{60}$.

42. Show that the formula

$$y_{n+1} = y_n[p + 1 - ay_n^p]/p$$

gives a second-order iterative process for $1/a^{1/p}$.

43. Devise a second-order iterative method on the lines of Example 42 to find $(1/9a)^{1/9}$. Use it to find $(8/9)^{1/9}$ and $(1/9)^{1/9}$ to eight decimals; check by verifying that the ratio of the results is $2^{1/3}$.

BIBLIOGRAPHY

THIS bibliography includes, as suggestions for further reading, some books and papers not referred to in the text. Page numbers in italics at the end of an entry indicate the reference in the text to the book or paper listed.

- AITKEN, A. C., 'On interpolation by proportional parts, without the use of differences', *Proc. Edin. Math. Soc.* (2), **3** (1932), 56—83.
- 'Studies in practical mathematics, II. The evaluation of the latent roots and the latent vectors of a matrix', *Proc. Roy. Soc. Edin.* **57** (1937), 269.
- 'Studies in practical mathematics, III. The application of quadratic extrapolation to the evaluation of a derivative and to inverse interpolation', *ibid.* **58** (1938), 161.
- 'Studies in practical mathematics, V. On the iterative solution of a system of linear equations', *ibid.* **63 A** (1950), 52.
- 'Studies in practical mathematics, VI. On the factorization of polynomials by iterative methods', *ibid.* **63 A** (1951), 174—224.
- ALLEN, D. N. DE G., *Relaxation Methods* (McGraw-Hill, 1954).
- Barlow's *Tables of Squares, Cubes, etc.* (edited by L. J. COMRIE, fourth edition, Spon, 1941)—20, 213.
- BICKLEY, W. G., 'Difference and associated operators, with some applications', *Journ. of Math. and Phys.* **27** (1948), 183—55.
- 'Finite difference formulae for the square lattice', *Quart. J. Mech. and Applied Math.* **1** (1948), 35—237.
- BICKLEY, W. G., and MILLER, J. C. P., 'The numerical summation of slowly convergent series of positive terms', *Phil. Mag.* (7), **22** (1936), 754—268.
- BICKLEY, W. G. See TEMPLE and BICKLEY.
- BIRKHOFF, G. D., and YOUNG, D. M., 'Numerical quadrature of analytic and harmonic functions', *Journ. of Math. and Phys.* **29** (1950), 217—240.
- BOOTH, A. D., *Numerical Methods* (Butterworth, 1955).
- BOOTH, A. D., and BOOTH, K. V. H., *Automatic Digital Calculators* (Butterworth; second edition, 1956)—284.
- British Association Mathematical Tables*, Part-volume B, *The Airy Integral* (1946)—21, 77, 128, 149.
- BROMWICH, T. J. I'A., *Theory of Infinite Series* (Macmillan, second edition, 1926)—64.
- Chambers's Six-figure Mathematical Tables* (edited by L. J. COMRIE), vol. 2 (1949)—20, 46, 64, 68, 69, 70, 71, 74, 75, 76, 83, 91, 114, 215, 235.
- Chambers's Shorter Six-figure Mathematical Tables* (edited by L. J. COMRIE) (1950)—20, 114.
- CHANDRASEKHAR, S., 'On the radiative equilibrium of a stellar atmosphere', *Astrophys. Journ.* **100** (1944), 76—123.
- *Radiative Transfer* (Clarendon Press, 1950)—123.
- CHERRY, T. M., 'Summation of slowly convergent series', *Proc. Camb. Phil. Soc.* **46** (1950), 436—268.
- COLLATZ, L., *Numerische Behandlung von Differentialgleichungen* (Springer, Second edition, 1955).
- *Eigenwertprobleme und ihre numerische Behandlung* (Chelsea, 1948).

- COMRIE, L. J., 'On the construction of tables by interpolation', *Month. Notices, Royal Astron. Soc.* **88** (1928), 506—79.
- 'Inverse interpolation and scientific applications of the National accounting machine', *Journ. Roy. Stat. Soc.*, supplement, **3** (1936), 87—25, 82.
- See *Chambers's Six-figure Mathematical Tables*.
- CRANK, J., *The Mathematics of Diffusion* (Clarendon Press, 1956) —254.
- CRANK, J., and NICOLSON, P., 'A practical method for numerical evaluation of partial differential equations of the heat conduction type', *Proc. Camb. Phil. Soc.* **43** (1947), 50—256.
- CROUT, P. D., 'A short method of evaluating determinants and solving sets of linear equations with real or complex coefficients', *Trans. Amer. Inst. Elect. Eng.* **60** (1941), 1235—180.
- DIJKSTRA, E. W., and VAN WIJNGAARDEN, A., *Table of Everett's Interpolation Coefficients* (Mathematisch Centrum, Amsterdam, 1955)—70.
- DOODSON, A. T., 'A method for the smoothing of numerical tables', *Quart. J. Mech. and Applied Math.* **3** (1950), 217—276.
- DWYER, P. S., *Linear Computations* (John Wiley, 1951).
- ECKERT, W. J., *Punched Card Methods in Scientific Computation* (Columbia Univ., 1940), 25.
- ECKERT, W. J., and JAMES, R., *Faster, faster* (McGraw Hill, 1955) —284.
- EYRES, N. R., *et al.* 'The calculation of variable heat flow in solids', *Phil. Trans. Roy. Soc.* **240** (1946), 1—254.
- FERRAR, W. L., 'On the Cardinal Function of Interpolation Theory', *Proc. Roy. Soc. Edin.* **45** (1925), 269; **46** (1926), 323—93. 'On the consistency of Cardinal Function Interpolation, *ibid.* **47** (1927), 230—93.
- FLETCHER, A., MILLER, J. C. P., and ROSENHEAD, L., *Index of Mathematical Tables* (Scientific Computing Service, 1946)—21, 69, 74, 220.
- FOX, L., 'Some improvements in the use of relaxation methods for the solution of ordinary and partial differential equations', *Proc. Roy. Soc. A*, **190** (1947), 31—245.
- 'A short summary of relaxation methods', *Quart. J. Mech. and Applied Math.* **1** (1948), 253.
- 'The solution by relaxation methods of ordinary differential equations', *Proc. Camb. Phil. Soc.* **45** (1949), 50—190, 196, 209.
- 'Practical methods for the solution of linear equations and the inversion of matrices', *Journ. Roy. Stat. Soc. B*, **12** (1950), 120—184.
- *The Numerical Solution of two-point Boundary Problems in Ordinary Differential Equations* (Clarendon Press, 1957)—159, 191, 196.
- see also NATIONAL PHYSICAL LABORATORY.
- FOX, L., and GOODWIN, E. T., 'Some new methods for the integration of ordinary differential equations', *Proc. Camb. Phil. Soc.* **45** (1949), 373—147, 154.
- 'The numerical solution of non-singular linear integral equations', *Phil. Trans. Roy. Soc.* **245** (1953), 501.
- FOX, L., HUSKEY, H. D., and WILKINSON, J. H., 'Notes on the solution of algebraic linear simultaneous equations', *Quart. J. Mech. and Applied Math.* **1** (1948), 149—180.
- FOX, L., and ROBERTSON, H. H. See NATIONAL PHYSICAL LABORATORY.
- FRIEDMAN, B., 'Note on approximating complex zeros of a polynomial', *Commun. on Pure and Applied Mathematics*, **2** (1949), 195—224.

- GOODWIN, E. T., 'The evaluation of integrals of the form $\int_{-\infty}^{\infty} f(x)e^{-x^2} dx$ ', *Proc. Camb. Phil. Soc.* **45** (1949), 241—117.
- GOODWIN, E. T., and STATON, J., 'Table of $\int_0^{\infty} [e^{-u^2}/(u+x)] du$ ', *Quart. J. Mech. and Applied Math.* **1** (1948), 319—119.
- HARTREE, D. R., 'Notes on iterative processes', *Proc. Camb. Phil. Soc.* **45** (1948), 230—217.
- 'A solution of the laminar boundary layer equation for retarded flow', *Aero. Res. Comm., Rep. and Mem.*, No. 2426 (1939, issued 1949)—256.
- HARTREE, D. R., KRONIG, R. DE L., and PEDERSEN, H., 'A theoretical calculation of the fine structure for the K -absorption band of Ge in GeCl_4 ', *Physica*, **1** (1934), 895—155.
- HARTREE, D. R., and WOMERSLEY, J. R., 'A method for the numerical or mechanical solution of certain types of partial differential equations', *Proc. Roy. Soc.* **161** (1937), 363—255.
- HILDEBRAND, F. B., *Introduction to Numerical Analysis* (McGraw Hill, 1956).
- HOUSEHOLDER, A. S., *Principles of Numerical Analysis* (McGraw-Hill, 1953).
- Index of Mathematical Tables.* See FLETCHER, A.
- INSTITUTION OF ELECTRICAL ENGINEERS. Convention on Digital Computer Techniques, *Proc. I.E.E.* **103** (1956). Part B, Suppl. No. 1—142, 206.
- Interpolation and Allied Tables* (H.M.S.O. 1956)—20, 21, 62, 66, 69, 70.
- KOPAL, Z. *Numerical Analysis* (Chapman and Hall, 1955)—122, 124, 127.
- LANCZOS, C., 'Trigonometric interpolation of empirical and analytic functions', *Journ. of Math. and Phys.* **17** (1938), 123.
- 'An iteration method for the solution of the eigenvalue problem of linear differential and integral operators', *Journ. of Research, Nat. Bur. Standards*, **45** (1950), 255.
- 'Spectroscopic eigenvalue analysis', *Journ. Washington Acad. Sci.* **45** (1955), 315.
- LAX, P., 'Weak solutions of non-linear hyperbolic equations and their numerical computation', *Commun. on Pure and Appl. Math.*, **7** (1954), 159.
- LEIGH, D. C. F., 'The laminar boundary layer; a method of solution by means of an automatic computer', *Proc. Camb. Phil. Soc.* **51** (1955), 320—256.
- LIEBMANN, H., 'Die angenäherte Ermittlung harmonischer Funktionen und konformer Abbildung', *Sitz. Bayer. Akad. München* (1918), 385—253.
- LOWAN, A. N., DAVIDS, N., and LEVENSON, A., 'Table of the zeros of the Legendre polynomials of order 1–16 and the weights in Gauss' mechanical quadrature formula', *Bull. Amer. Math. Soc.* **48** (1942), 739—122.
- MACFARLANE, G. G., 'The application of Mellin transforms to the summation of slowly convergent series', *Phil. Mag.* (7), **40** (1949), 188—268.
- MADLUNG, E., 'Über eine Methode zur schnellen numerischen Lösung von Differentialgleichungen zweiter Ordnung', *Zeit. f. Phys.* **67** (1931), 516—154.
- MANNING, M. F., and MILLMAN, J., 'Note on numerical integration', *Phys. Rev.* **53** (1938), 673—142.
- MICHEL, J. G. L., 'Central difference formulae obtained by means of operator expansions', *Journ. Inst. of Actuaries*, **72** (1946), 470—64.

- MILLER, J. C. P., 'Checking by differences', *Math. Tables and Aids to Comp.* **4** (1950), 3—48.
- 'A method for the determination of converging factors, applied to the asymptotic expansions of the parabolic cylinder functions', *Proc. Camb. Phil. Soc.* **48** (1952), 243.
- See also *Brit. Assn. Math. Tables*, BICKLEY, and FLETCHER.
- MILNE, W. E., *Numerical Calculus* (Univ. of Princeton Press, 1950)—129.
- 'Numerical determination of characteristic numbers', *Journ. of Research, Nat. Bur. Standards*, **45** (1950), 245—165.
- MILNE, W. E., *Numerical Integration of Differential Equations* (Wiley, 1953).
- MINEUR, H., *Technique de Calcul Numérique* (Ch. Béranger, 1952)
- MORRIS, J., 'An escalator process for the solution of linear simultaneous equations', *Phil. Mag.* (7), **37** (1946), 106—170.
- NATIONAL PHYSICAL LABORATORY. *Proceedings of a Symposium on Automatic Digital Computation*, March 1953 (H.M.S.O., 1954), especially: Paper 18. J. H. Wilkinson, 'Linear algebra on the pilot ACE'; Paper 19. L. Fox and H. H. Robertson, 'The numerical solution of differential equations'—191, 206.
- *Mathematical Tables Vol. I; The Use and Construction of Mathematical Tables*, by L. Fox (H.M.S.O., 1956)—20, 61, 66.
- NEVILLE, E. H., 'Iterative interpolation', *Journ. Indian Math. Soc.* **20** (1934), 87—85.
- OLVER, F. W. J., 'A new method for the evaluation of zeros of Bessel functions and of other solutions of second-order differential equations', *Proc. Camb. Phil. Soc.* **46** (1950), 570—142.
- 'The evaluation of zeros of high-degree polynomials', *Phil. Trans. Roy. Soc.* **244** (1952), 385—221.
- RICHARDSON, L. F., 'The approximate arithmetical solution by finite differences of physical problems involving differential equations', *ibid.* **210** (1910), 307—153.
- 'The deferred approach to the limit', *ibid.* **226** (1927), 300—134, 152.
- 'A purification method for computing the latent columns of numerical matrices and some integrals of differential equations', *ibid.* **242** (1950), 439—201.
- RIDLEY, E. C., 'The self-consistent field for Mo⁺', *Proc. Camb. Phil. Soc.* **51** (1955), 702—163.
- 'A numerical method for solving second-order linear differential equations with two-point boundary conditions', *Proc. Camb. Phil. Soc.* **53** (1957), 442—162.
- SALZER, H. E., and ZUCKER, R., 'Table of the zeros and weight factors of the first 15 Laguerre polynomials', *Bull. Amer. Math. Soc.* **55** (1949), 1004—124.
- SAMUELSON, P. A., 'Iterative computation of complex roots', *Journ. of Math. and Phys.* **28** (1949), 259—222.
- SOUTHWELL, R. V., 'Stress calculations in frameworks by the method of "systematic relaxation of constraints"', *Proc. Roy. Soc. A*, **151** (1935), 56—185, 190.
- 'On relaxation methods; a mathematics for engineering science' (Bakerian Lecture), *ibid.* **184** (1945), 253—185.
- *Relaxation Methods in Engineering Science* (Oxford, 1940)—185.
- *Relaxation Methods in Theoretical Physics* (Oxford, 1946)—251.

- SPENCER, J., 'On the graduation of the rate of sickness and mortality presented by the experience of the Manchester Unity of Oddfellows during the period 1893-97', *J. Inst. Actuaries*, **38** (1904), 334-275.
- STEFFENSEN, J. F., *Interpolation* (Chelsea, 1950).
- STIEFEL, L., 'Über einige Methoden der Relaxationsrechnung', *Zeit. f. angew. Math. und Phys.* **3** (1952), 1-251.
- SZÁSZ, O., 'Summation of slowly convergent series', *Journ. of Math. and Phys.* **28** (1949), 272-268.
- TAUSSKY, O., 'Note on the condition of matrices', *Math. Tables and Aids to Comp.* **4** (1950), 111-171.
- TEMPLE, G., 'The general theory of relaxation methods applied to linear systems', *Proc. Roy. Soc. A*, **169** (1938), 476-189.
- TEMPLE, G., and BICKLEY, W. G., *Rayleigh's Principle* (Oxford, 1933)-165, 199.
- THOMAS, L. H., 'Computation of one-dimensional compressible flow, including shocks', *Commun. on Pure and Appl. Math.* **7** (1954), 159-259.
- TURING, A. M., 'Rounding-off errors in matrix processes', *Quart. J. Mech. and Applied Math.* **1** (1948), 287-169, 180.
- WHITEHEAD, A. N., *Introduction to Mathematics* (Home University Library Series, No. 18; Oxford, 1948)-2.
- WHITTAKER, E. T., 'On the functions which are represented by the expansions of interpolation theory', *Proc. Roy. Soc. Edin.* **35** (1915), 181-93.
- WHITTAKER, E. T., and ROBINSON, G., *Calculus of Observations* (Blackie, fourth edition, 1944)-122, 221, 275.
- WHITTAKER, E. T., and WATSON, G. N., *Modern Analysis* (Camb. Univ. Press, fourth edition, 1927)-101, 271, 288.
- WICK, G. C., 'Über ebene Diffusionsprobleme', *Zeit. f. Phys.* **121** (1943), 702-123.
- WILKES, M. V., 'A method of solving second-order simultaneous linear differential equations using the Mallock machine', *Proc. Camb. Phil. Soc.* **36** (1940), 204-142.
- *Automatic Digital Computers* (Methuen, 1956)-284.
- 'Solution of linear algebraic and differential equations by the long division algorithm', *Proc. Camb. Phil. Soc.* **52** (1956), 758-180.
- WILKES, M. V., WHEELER, D. J., and GILL, S., *Preparation of Programs for an Electronic Digital Computer* (Addison-Wesley Press, 1951, second edition in preparation)-284.
- WILKINSON, J. H., 'The calculation of latent roots and vectors of matrices on the pilot model of the ACE', *Proc. Camb. Phil. Soc.* **50** (1954), 538-206.
- 'An assessment of optimum coding using the pilot model ACE', *Phil. Trans. Roy. Soc.* **248** (1955), 243-284.
- see also NATIONAL PHYSICAL LABORATORY.
- WOODWARD, P. M., 'Tables of interpolation coefficients for use in the complex plane', *Phil. Mag.* (7), **39** (1948), 594-240.
- WOODWARD, P. M., and WOODWARD, A. M., 'Four-figure tables of the Airy function in the complex plane', *ibid.* **37** (1946), 236-240.

INDEX

- Accumulator, of desk machine, 11, 15, 48, 60; of automatic digital machine, 282.
- Accuracy, 9, 190, 273.
- Adding machine, 11, 19, 42.
- Addition, on desk machine, 12, 14.
- of complex numbers, 235.
- Airy functions, 77, 128, 264.
- AITKEN, A. C., 83.
- Algebraic equations: linear simultaneous, 166–209; non-linear, in one variable, 210–27; non-linear, in two or more variables, 228–34.
- Analysis of observations, 4, 268.
- Automatic digital calculating machines, 25, 279–86; accumulator of, 282; arithmetical unit of, 282; control in, 280, 282; instruction in, 281, 282; program for, 281; serial storage in, 282; store of, 280; sub-routine for, 280, 283; word in, 281.
- Automatic division, on desk machine, 19.
- Auxiliary function in smoothing, 276.
- variables in tables, 22.
- Averaging operator μ , 38, 52; inverse of, 62.

- Back-substitution, 175, 178.
- Bernoulli numbers, 101.
- Bessel functions, 23, 27, 190, 271.
- interpolation coefficients, 68, 69, 95.
- Bessel's interpolation formula, 64, 67–73, 86, 90, 91, 95, 98, 117.
- BICKLEY, W. G., 55, 91, 199, 237.
- Boundary conditions, 134, 155, 240, 243.

- Calculating machines: automatic, 279–86; desk, 11–20, 27.
- Cardinal function, 93–96, 117.
- Characteristic values: of matrices, 169, 171, 196–209; of ordinary differential equations, 162–5, 209; of partial differential equations, 252.
- vectors of matrices, 196–208.
- Characteristics, of hyperbolic partial differential equations, 257–63.
- Checking and checks, 3, 4, 17, 28, 29, 30, 39, 42, 43, 44, 47, 62, 74, 81, 82, 83, 93, 104, 105, 106, 108, 136, 137, 138, 140, 151, 160, 173, 174, 184, 187, 189, 201, 205, 222, 230, 236, 265, 267, 274.
- Choleski method, for linear simultaneous algebraic equations, 180; for matrix inversion, 185; Thomas-Fox application to differential equations, 191.
- Circular functions, 32, 34, 60, 62, 78, 286.
- Clearing, 11.
- Complement, 15.
- Complex numbers, 235, 236.
- variable, functions of, 235, 239.
- COMRIE, L. J., 64, 68, 70, 75, 76, 79, 82, 83, 91.
- Continued product, 27.
- Convergence, 3, 189.
- Critical tables, 21.
- Cross-sum check, 173, 174, 176, 177, 179.
- Cubic equation, 220.
- Cylindrical coordinates, finite differences in, 240.

- Dedekind section, 3.
- Deferred approach to the limit, 151.
- Derivatives and differences, 49, 50, 55–58, 126, 127.
- Determinant, 166, 168, 169, 177, 197.
- Difference operators, 36, 50–58; inverses of 52, 62.
- Differences, finite, 35–58; building up from, 41, 48; checking by, 39, 43; and derivatives, 37, 49, 55–58, 191, 193; effects of errors on, 36, 43–46; in terms of function values, 38, 53; notation for, 37; smoothing by, 277.
- Differentiation: formulae, 55, 126–7; graphical, 129; numerical, 124–8.
- Diffusion equation, 242, 254–7.
- Digital machine, 279.
- Direct interpolation, 59.
- Divided differences, 86–89.
- Division, on desk machine, 17.
- of complex numbers, 235.
- of a polynomial by a quadratic, 222.
- DOODSON, A. T., 276.

- Elimination, 167, 173–9, 228.
- Elliptic partial differential equations, 190, 242, 243, 244–53.
- End-figure method, for subtabulation, 79.
- Errors: *see* Rounding errors, Systematic errors, Truncation errors, Mistakes.
- Euler-Maclaurin formula, 101, 104, 115, 116, 117, 266, 269.
- Euler's transformation of a power series, 265.
- Everett interpolation coefficients, 67, 70.
- Everett's interpolation formulae, 64, 66, 70–74, 75, 80, 86, 98.
- Exponential extrapolation, 30, 90, 216.
- function, 27, 76, 286.

- Factorial polynomials, 40, 65, 66, 67, 68.
 Factorization: of differential equation, 161;
 of matrix into triangular matrices, 180,
 192.
 FERRAR, W. L., 93, 95.
 Finite differences: in cylindrical coordi-
 nates, 240; in two dimensions, 236; *see*
 also Differences, finite.
 Formulae; evaluation of, 4, 26–32; sig-
 nificance of, in numerical work, 26.
 Forward differences, 37, 51, 114, 265; in
 integration, 114; in interpolation, 63.
 Fox, L., 147, 154, 190, 191, 196, 245.
 Frequency analysis, 4.
 Function, continuous, 33.
 — of a complex variable, 235, 239.
 — of two variables, 235–63.

 Gamma function, 33.
 Gauss integration formulae, 120–4, 286.
 Gibbs phenomenon, 268.
 GOODWIN, E. T., 117, 119, 147, 154.
 Graphical methods, 129, 210, 228, 230–3,
 276.
 Gregory integration formula, 114.
 Group relaxations, 188.
 Guarding figures, 6, 174; in interpolation,
 71, 72, 81.

 h^2 -extrapolation, 148, 152, 254, 255, 274.
 Half-way interpolation, 61, 108.
 Harmonic analysis, 4, 94, 268.
 Heat-conduction equation, 242, 254–7.
 Hollerith machines, 25.
 Hyperbolic functions, 27.
 — partial differential equations, 243,
 253–8.

 Ill-conditioned equations, 168, 177, 179,
 190, 206.
 Indeterminate forms, 30.
 Initial conditions, 24, 243.
 Input, in automatic machine, 280.
 Integral; between fixed limits, 97, 113–24;
 by solution of a differential equation,
 119; as function of upper limit, 97,
 104–9.
 — condition, on solution of differential
 equation, 135, 165.
 — equation, 4.
 — parametric, 118.
 — twofold, 112.
 Integrating factor, 146.
 Integration formula, 57, 58, 98–104.
 Integration, numerical, 4; of a given func-
 tion of x , 97–123; of an ordinary
 differential equation, 97, 134–65, 191–6,
 of a partial differential equation, 242–63.

 Interpolation, 4, 22, 59–96; linear, 59, 60;
 non-linear, 23, 59, 61–96; in complex
 plane, 240.
 Interval length, change of: in quadrature,
 108; in integration of differential equa-
 tion, 139, 249.
 Inverse interpolation, 59, 60, 70, 75, 89,
 210, 215, 229.
 — of a matrix, 7, 168, 178, 185, 191, 199.
 — operators, 52, 55, 62.
 Iterative process, 211, 285, 292; for alge-
 braic equation, 211; for characteristic
 values of a matrix, 199; for differential
 equations, 153, 157, 193; for inverse
 interpolation, 90, 214; for quadratic
 factor of polynomial, 224.

 Jury problem, 135.

 Lagrange integration formula, 114.
 — interpolation coefficients, 74.
 — interpolation formula; equal intervals
 of argument, 74–76; for inverse inter-
 polation, 91–93; unequal intervals of
 argument, 82, 86, 114, 122.
 Laplace's equation, 235, 239, 246, 248, 250,
 253, 269.
 Laplacian operator; in two dimensions,
 235, 238; in three dimensions, 240.
 Latent roots (of matrices), *see* Characteris-
 tic values.
 — vectors (of matrices), *see* Characteristic
 vectors.
 Leading differences, 37.
 Legendre polynomials, 120, 271.
 LIEBMANN, H., 253.
 Liebmann's process for Laplace's equation,
 253.
 Limiting process, 3.
 Linear cross-mean, 84.
 — differential equations; *see* Ordinary,
 Partial differential equations.
 — interpolation, 59, 60.
 — operators, 52, 93, 130.
 — simultaneous equations, *see* Simulta-
 neous equations.
 Lower triangular matrix, 168, 180, 185,
 192.

 MADELUNG, E., 154.
 Madelung transformation, 154.
 Marching problem, 134.
 Matching process, in solution of linear
 differential equation with two-point
 boundary conditions, 160, 163, 164.
 Matrices, 4, 168, 171, 180; characteristic
 values of, 171, 196–209; inverse and in-
 version of, 168, 178, 185, 191, 199.

- Mean differences, 38.
 MICHEL, J. G. L., 64.
 MILLER, J. C. P., 47.
 MILNE, W. E., 93, 129.
 Mistakes, 5, 7, 27, 44, 45, 105, 174, 215.
 Modified differences, 70, 75.
 MORRIS, J., 170.
 Multiple roots of algebraic equations, 217, 222.
 Multiplication, on desk machine, 12, 16.
 — of complex numbers, 235.
 'National' calculating machine, 24, 77.
 Neighbouring roots of algebraic equation, 217, 218.
 Newton's forward-difference interpolation formula, 63.
 — iterative process for square root, 213.
 Newton-Raphson iterative process, 214, 215.
 Nominal accuracy, *see* Precision.
 Non-linear interpolation, 59, 61–91.
 Normal equations, 171.
 Normalization of characteristic vectors, 197, 199.
 — condition on solution of a differential equation, 165.
 NUMEROV, B., 142.
 OLVER, F. W. J., 142.
 Operations table, in relaxation calculation, 186, 189, 208.
 Order of an iterative process, 212.
 Ordinary differential equations, 2, 24, 119, 134–65, 191, 255, 285; boundary conditions for, 134, 155; first-order, 143, 146; linear, 23, 135, 142, 148, 154, 155; second-order, with first derivative absent, 135–43, 191, with first derivative present, 148, with two-point boundary conditions, 159–65; second-order linear, relation to simultaneous algebraic equations, 191–6; third and higher orders, 149.
 Orthogonality of characteristic vectors of a symmetrical matrix, 197, 200, 201, 205, 206; of Legendre polynomials, 121.
 Parabolic partial differential equations, 243, 253–7, 258.
 Partial differential equations, 242; boundary conditions for, 243; elliptic, 190, 243, 244–53, 258; hyperbolic, 243, 257–62; parabolic, 243, 253–7, 258.
 — fractions, 83.
 Pivotal coefficient, 174, 177.
 — equation, 173, 175, 177.
 — value, 77, 78, 79, 80.
 Poisson's equation, 242, 244, 248, 249.
 Polar coordinates, 235, 240.
 Polynomial; differences of, 39; divided differences of, 87; derivatives of, 88; evaluation of, 28, 41; factorial, 40.
 — equation, 218, 221.
 Power series, 29, 32, 264.
 Powers-Samas machines, 25.
 Precision, 9.
 Processes, numerical, 1, 3, 26.
 Proportional parts, in interpolation, 61.
 Punched-card machines, 25.
 Purification process, for characteristic vectors of matrices, 201.
 Quadratic equation, 32, 219.
 — factor of polynomial, 224.
 Quadrature, 97–124.
 Quartic equation, 220.
 Rayleigh's principle, 199.
 Recurrence relation, 27, 190, 271.
 Reduced derivatives, 76, 150.
 Relaxation method; for algebraic equations, 185–91, 257, 285; for characteristic values of matrices, 207; for ordinary differential equations with two-point boundary conditions, 196; for elliptic partial differential equations, 245–53.
 Residuals, 167, 171, 177, 179, 186, 187, 188, 245, 248.
 Resistance function, 275.
 Riccati transformation, 155, 201.
 RICHARDSON, L. F., 134, 152, 201, 253.
 Richardson-Liebmann process, 253, 256.
 Root-squaring process, 221.
 Rounding errors, 5, 45, 46, 106, 141, 177, 179, 217, 218, 258, 272; *see also* Tolerance.
 Round-off, 5, 42, 43.
 Rule of false position, 214.
 Second difference, 35, 127; direct evaluation from function values, 47; building up from, 48, 79, 80, 136; *see also* Differences, finite.
 Second-order iterative processes, 212, 216, 226.
 Separation of variables, 242.
 Series, evaluation of, 29, 32, 265.
 Shift operators E , E^{-1} , 51, 69, 265.
 Short-cutting, 17, 18.
 Simpson's rule, 50, 102, 106, 115, 240, 285; correction to, 102, 240.
 Simultaneous algebraic equations; linear, 4, 166–94, 244, 257; non-linear, 4, 228–34.
 Singularity, 22, 110, 118, 241; integration in neighbourhood of, 110, 118.

- Slide-rule, 11, 23, 28.
 Smoothing, 4, 39, 272-8.
 Smoothness, 34, 47, 272.
 SOUTHWELL, R. V., 185, 188, 237, 251.
 SPENCER, J., 275.
 Square root, 213; of complex numbers, 236.
 Statistics, 3, 4.
 STATON, J., 119.
 Sub-tabulation, 4, 70, 77-82.
 Subtraction, on desk machine, 12, 14.
 Summation of series, *see* Series.
 — operator σ , 52, 104, 277.
 Symmetrical matrix, 170, 197-201.
 Systematic error, 43, 44, 78.
- Tables, mathematical, 11, 20, 69, 74.
 — critical, 21.
 Tabulation, 1, 4; use of auxiliary variables in, 22.
 Taylor series, 49, 51, 53, 63, 76, 149; in interpolation, 76; in integration of differential equations, 149.
 TEMPLE, G., 189, 199.
 Terminal conditions, 243.
 Throw-back, 70, 71, 72, 75, 90.
 Tolerance, for rounding errors, 5, 6, 106, 113, 128, 205, 219.
- Total, on adding machine, 20.
 Transfer, on desk machine, 15; on automatic machine, 280.
 Trapezium rule, for integration, 99, 103.
 Triangular matrix, 168, 175, 180-5, 192-3.
 Truncation errors, 5, 93, 129, 141, 244, 255, 274.
 Two dimensions; finite differences in, 236-42; Laplace's equation in, 235, 239, 244, 245-51, 253.
 Two-point boundary conditions in differential equations, 34, 155-65, 191, 255, 273; in linear equations, 159-65, 191-6.
 TURING, A. M., 169.
- Upper triangular matrix, 168, 175, 180, 192.
- Wave equation, 242.
 Weber functions, 271.
 Weddle's rule, for quadrature, 103, 115.
 WHITTAKER, E. T., 93.
 Whittaker functions, 60, 271.
 WILKES, M. V., 180.
 WILSON, T. S., 170.
 Working sheet, 8, 25, 26, 275, 279, 280.

PRINTED IN GREAT BRITAIN
AT THE UNIVERSITY PRESS, OXFORD
BY VIVIAN RIDLER
PRINTER TO THE UNIVERSITY

DATE DUE

| | | |
|-------------|-------|-------------------|
| JAN 15 1988 | APR 1 | 2003 |
| JAN 15 1989 | | |
| SEP 15 1989 | | |
| SEP 15 1992 | | |
| AUG 31 1992 | | |
| APR 10 2003 | | |
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |
| 201-6503 | | Printed in USA |

TRENT UNIVERSITY



0 1164 0015984 8

004780

QA Hartree, Douglas Rayner
297 Numerical analysis. 2d ed.
H3
1961
Trent
University

